

## ABSTRACT

Title of dissertation:      FEATURE LEARNING AND  
                                     ACTIVE LEARNING FOR  
                                     IMAGE QUALITY ASSESSMENT

Peng Ye, Doctor of Philosophy, 2014

Dissertation directed by:   Professor Rama Chellappa  
                                     Department of Electrical and  
                                     Computer Engineering

Dr. David Doermann  
University of Maryland Institute for  
Advanced Computer Studies

With the increasing popularity of mobile imaging devices, digital images have become an important vehicle for representing and communicating information. Unfortunately, digital images may be degraded at various stages of their life cycle. These degradations may lead to the loss of visual information, resulting in an unsatisfactory experience for human viewers and difficulties for image processing and analysis at subsequent stages. The problem of visual information quality assessment plays an important role in numerous image/video processing and computer vision applications, including image compression, image transmission and image retrieval, etc. There are two divisions of Image Quality Assessment (IQA) research – Objective IQA and Subjective IQA. For objective IQA, the goal is to develop a computational model that can predict the quality of distorted image with respect to human perception or other measures of interest accurately and automatically.

For subjective IQA, the goal is to design experiments for acquiring human subjects' opinions on image quality. It is often used to construct image quality datasets and provide the groundtruth for building and evaluating objective quality measures. In the thesis, we will address these two aspects of IQA problem.

For objective IQA, our work focuses on the most challenging category of objective IQA tasks - general-purpose No-Reference IQA (NR-IQA), where the goal is to evaluate the quality of digital images without access to reference images and without prior knowledge of the types of distortions.

First, we introduce a feature learning framework for NR-IQA. Our method learns discriminative visual features in the spatial domain instead of using hand-craft features. It can therefore significantly reduce the feature computation time compared to previous state-of-the-art approaches while achieving state-of-the-art performance in prediction accuracy.

Second, we present an effective method for extending existing NR-IQA models to "Opinion-Free" (OF) models which do not require human opinion scores for training. In particular, we accomplish this by using Full-Reference (FR) IQA measures to train NR-IQA models. Unsupervised rank aggregation is applied to combine different FR measures to generate a synthetic score, which serves as a better "gold standard". Our method significantly outperforms previous OF-NRIQA methods and is comparable to state-of-the-art NR-IQA methods trained on human opinion scores.

Unlike objective IQA, subjective IQA tests ask humans to evaluate image quality and are generally considered as the most reliable way to evaluate the visual

quality of digital images perceived by the end user. We present a hybrid subjective test which combines Absolute Categorical Rating (ACR) tests and Paired Comparison (PC) tests via a unified probabilistic model and an active sampling method. Our method actively constructs a set of queries consisting of ACR and PC tests based on the expected information gain provided by each test and can effectively reduce the number of tests required for achieving a target accuracy. Our method can be used in conventional laboratory studies as well as crowdsourcing experiments. Experimental results show our method outperforms state-of-the-art subjective IQA tests in a crowdsourced setting.

# FEATURE LEARNING AND ACTIVE LEARNING FOR IMAGE QUALITY ASSESSMENT

by

Peng Ye

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2014

Advisory Committee:  
Professor Rama Chellappa, Chair  
Dr. David Doermann, Co-Chair  
Professor Larry Davis  
Professor Min Wu  
Professor Ramani Duraiswami



© Copyright by  
Peng Ye  
2014

## Dedication

This dissertation is dedicated to my parents and my husband.

## Acknowledgments

First, I would like to thank my PhD advisors Dr. David Doermann and Professor Rama Chellappa, for supporting me during the past four and half years. Dr. Doermann has been extremely supportive and helpful throughout my entire doctoral journey. He always made himself available whenever I need help and patiently provided the vision, encouragement and ideas necessary for me to proceed. The completion of this dissertation would not have been possible without him. Professor Chellappa is one of the smartest and most knowledgeable people I know. His class on statistical pattern recognition is my first encounter with machine learning, which is crucially important for my research.

I would like to thank my committee members, Professors Min Wu, Larry Davis and Ramani Duraiswami for being on the committee, reviewing the dissertation and taking time out of their busy schedules. I am especially grateful to Professor Wu for being a role model of female researcher.

I would also like to thank all LAMP members: Dave, Wael, Elena, Jayant, Le, XianZhi, Rajiv, Chung-Ta, Jingtao, Sungmin and Varun. Jayant and Le have worked closely with me in the CORNIA project. I learned a lot from them and really enjoyed working with them.

I owe my deepest thanks to my parents for bringing me to the world, providing me good education and encouragement to pursue the doctoral degree. Finally, I thank my husband Kang Tu for supporting me through all the ups and downs.

# Table of Contents

List of Tables	vii
List of Figures	ix
List of Abbreviations	xi
1 Introduction	1
1.1 Feature Learning for No-Reference Image Quality Assessment . . . . .	8
1.2 Blind Learning of Image Quality based on Synthetic Scores . . . . .	11
1.3 Active Learning for Subjective Image Quality Assessment . . . . .	13
1.4 Outline . . . . .	14
2 Feature Learning for No-Reference Image Quality Assessment	15
2.1 Introduction . . . . .	15
2.1.1 Related work . . . . .	17
2.1.2 Our approach . . . . .	19
2.2 Feature Learning for NR-IQA . . . . .	20
2.2.1 Feature extraction . . . . .	22
2.2.1.1 Local feature encoding . . . . .	22
2.2.1.2 Summarizing statistics . . . . .	24
2.2.2 Unsupervised Filter learning . . . . .	26
2.2.3 Supervised Filter Learning . . . . .	27
2.2.3.1 Problem formulation . . . . .	28
2.2.3.2 Optimizing B . . . . .	29
2.3 Experiments . . . . .	33
2.3.1 Protocol . . . . .	33
2.3.2 Experiments on Natural Scene Image . . . . .	35
2.3.2.1 Dataset . . . . .	35
2.3.2.2 Impact of unsupervised CORNIA parameters . . . . .	37
2.3.2.3 Evaluating unsupervised CORNIA . . . . .	39
2.3.2.4 Evaluating the supervised CORNIA . . . . .	43
2.3.2.5 Evaluating Speed . . . . .	50

2.3.3	Experiments on Document Images . . . . .	51
2.4	Discussion and Open Problems . . . . .	54
2.4.1	Discussion . . . . .	54
2.4.2	Open Problems . . . . .	55
3	Blind Learning of Image Quality based on Synthetic Scores . . . . .	58
3.1	Introduction . . . . .	58
3.2	Related Work . . . . .	60
3.2.1	OF-NRIQA Models . . . . .	60
3.2.2	Combining Multiple Full-Reference Measures . . . . .	61
3.3	Unsupervised FR measure combination . . . . .	62
3.3.1	Full-Reference Measures . . . . .	63
3.3.2	Combining Full-Reference Measures . . . . .	64
3.3.3	Computational Complexity . . . . .	66
3.3.4	Training . . . . .	67
3.4	Experiments . . . . .	67
3.4.1	Experimental Protocol . . . . .	67
3.4.2	Implementation Details . . . . .	69
3.4.3	Evaluation . . . . .	70
3.4.3.1	Comparison with FR and OF-NRIQA Algorithms . . . . .	70
3.4.3.2	Comparison with OA-NRIQA Algorithms . . . . .	71
3.4.3.3	Comparison of the combined synthetic score and FR measures . . . . .	75
3.4.3.4	Combining SSIM and WMSE . . . . .	79
3.5	Discussion and Extension . . . . .	81
3.5.1	Discussion . . . . .	81
3.5.2	Extension . . . . .	83
4	Active Learning for Subjective Image Quality Assessment . . . . .	84
4.1	Introduction . . . . .	84
4.2	Related Work . . . . .	86
4.2.1	Crowdsourcable QoE . . . . .	86
4.2.2	Preference Aggregation . . . . .	87
4.3	Combining Ratings and Paired Comparisons . . . . .	89
4.3.1	Mean Opinion Score Test . . . . .	89
4.3.2	Paired Comparison Test . . . . .	91
4.3.3	Posterior Probability of the Underlying Score . . . . .	92
4.4	Subjective Experimental Design based on Active Sampling . . . . .	95
4.4.1	Information Measure of Experiments . . . . .	95
4.4.2	Active Sampling . . . . .	97
4.5	Experiments . . . . .	98
4.5.1	Dataset . . . . .	98
4.5.2	Evaluation Measure . . . . .	100
4.5.3	GroundTruth . . . . .	102
4.5.4	Evaluation . . . . .	103

4.6	Discussion and Open Problems . . . . .	106
4.6.1	Discussion . . . . .	106
4.6.2	Open Problems . . . . .	108
5	Conclusions . . . . .	110
5.1	Feature Learning for No-Reference Image Quality Assessment . . . . .	110
5.2	“Opinion-free” No-Reference Image Quality Assessment . . . . .	111
5.3	Active Learning for Subjective Image Quality Assessment . . . . .	112
5.4	Publications . . . . .	112
5.4.1	Feature Learning for NR-IQA . . . . .	112
5.4.2	“Opinion-free” NR-IQA . . . . .	113
5.4.3	Active Learning for subjective IQA . . . . .	114
	Appendix A . . . . .	115
A.1	Explicit Differentiation . . . . .	116
A.2	Implicit Differentiation . . . . .	117
A.3	Computation Details . . . . .	118
A.3.1	Derivative of $R_m$ . . . . .	119
A.3.2	Derivative of $R_p$ . . . . .	120
A.3.3	Derivative of $\partial \log Pr(M s, \gamma)/\partial s$ . . . . .	121
	Bibliography . . . . .	122

## List of Tables

1.1	Examples of image degradations or image degradation sources. . . . .	4
1.2	Features for objective Natural Scene IQA. . . . .	9
1.3	Features for objective Document IQA. . . . .	10
2.1	Evaluation on the LIVE distorted images. ( <i>Italicized</i> algorithms are NR-IQA algorithms, others are FR-IQA algorithms.) . . . . .	40
2.2	Evaluation on the TID2008 dataset. ( <i>Italicized</i> algorithms are NR-IQA algorithms, others are FR-IQA algorithms.) . . . . .	40
2.3	Evaluation on the CSIQ dataset. ( <i>Italicized</i> algorithms are NR-IQA algorithms, others are FR-IQA algorithms.) . . . . .	41
2.4	Evaluation on the LIVEMD dataset. ( <i>Italicized</i> algorithms are NR-IQA algorithms, others are FR-IQA algorithms.) . . . . .	41
2.5	Results of the two sample T-test performed between SROCC values obtained by different measures on the LIVE dataset. . . . .	42
2.6	Results of the two sample T-test performed between SROCC values obtained by different measures on the TID2008 dataset. . . . .	42
2.7	Results of the two sample T-test performed between SROCC values obtained by different measures on the CSIQ dataset. . . . .	43
2.8	Train on LIVE and test on TID2008 . . . . .	43
2.9	Train on LIVE and test on CSIQ . . . . .	43
2.10	Comparing supervised and unsupervised CORNIA on the LIVE dataset. . . . .	44
2.11	Comparing supervised and unsupervised CORNIA on the TID2008 dataset. . . . .	46
2.12	Comparing supervised and unsupervised CORNIA on the CSIQ dataset. . . . .	46
2.13	Comparing supervised and unsupervised CORNIA on the LIVEMD dataset. . . . .	47
2.14	Comparing supervised and unsupervised CORNIA with RBF kernel on the TID2008 dataset. . . . .	48
2.15	Comparing supervised and unsupervised CORNIA with RBF kernel on the CSIQ dataset. . . . .	48
2.16	Comparing supervised and unsupervised CORNIA with RBF kernel on the LIVEMD dataset. . . . .	49

2.17	Comparing CORNIA with PSNR, SSIM and BRISQUE on TID2008, CSIQ and LIVEMD datasets. . . . .	49
2.18	Feature extraction time (in seconds). . . . .	51
2.19	Median LCC and SROCC with 1000 iterations of experiments on the SOC dataset. . . . .	54
3.1	Pair-wise SROCC between FR measures and RRFscore (Evaluated on LIVE). . . . .	65
3.2	Parameters used in our experiments. . . . .	70
3.3	Results on LIVE. . . . .	72
3.4	Results on CSIQ. . . . .	72
3.5	Results on TID2008. . . . .	73
3.6	Results of the two sample T-test performed between SROCC values obtained by different measures (Evaluated on LIVE). . . . .	73
3.7	Results of the two sample T-test performed between SROCC values obtained by different measures (Evaluated on TID2008). . . . .	73
3.8	Results of the two sample T-test performed between SROCC values obtained by different measures (Evaluated on CSIQ). . . . .	74
3.9	Standard deviation of SROCC and LCC for 1000 iterations of experiments on LIVE. . . . .	74
3.10	Train on LIVE and test on TID2008 . . . . .	75
3.11	Train on LIVE and test on CSIQ . . . . .	75
3.12	Test FR measures on LIVE (779 distorted images): ‘original’–correlation between original FR measures and DMOS; ‘SS’–correlation between synthetic scores and DMOS. . . . .	76
3.13	Test FR measures on TID2008 (1700 distorted images): ‘original’–correlation between original FR measures and DMOS; ‘SS’–correlation between synthetic scores and DMOS. . . . .	79
3.14	Test FR measures on LIVE (779 distorted images): ‘original’–original FR measures and DMOS; ‘logistic’–FR measures with nonlinear fitting; ‘SS’–combined synthetic scores. . . . .	80
3.15	Test FR measures on TID2008 (1700 distorted images): ‘original’–original FR measures and DMOS; ‘logistic’–FR measures with nonlinear fitting; ‘SS’–combined synthetic scores. . . . .	80
4.1	Correlation with LIVE’s DMOS . . . . .	103
4.2	Average number of required observations per image for achieving a given Kedall’s $\tau$ . . . . .	108



## List of Figures

1.1	Document image generation process. . . . .	3
1.2	Natural scene image generation process. . . . .	3
1.3	Examples of distorted document images. . . . .	5
1.4	Examples of distorted natural scene images (a) Undistorted reference image. (b) JPEG2000 Compression. (c) JPEG Compression. (d) White Gaussian Noise. (e) Gaussian Blur. (f) Fast Fading. . . . .	5
1.5	Objective IQA categories. . . . .	7
1.6	Overview of NR-IQA systems (a) Conventional NR-IQA systems. (b) Unsupervised CORNIA. (c) Supervised CORNIA. . . . .	12
2.1	Learning Features and Prediction Model for NR-IQA. . . . .	21
2.2	Examples of filter responses for different types and levels of distortions (High DMOS indicates low quality). . . . .	24
2.3	Randomly selected centroids trained on CSIQ database using K-means. . . . .	27
2.4	Overview of supervised filter learning method with linear SVR. . . . .	27
2.5	Effect of codebook size (tested on the LIVE database). . . . .	37
2.6	Effect of different encoders, evaluated using SROCC on the LIVE database. . . . .	38
2.7	Optimization process of the first 51 iterations on training set and validation set (average LCC from 100 fold experiments on LIVE). . . . .	45
2.8	(a) 200 $5 \times 5$ filters in CB200. (b) 200 $5 \times 5$ filters SF200. . . . .	50
2.9	Examples of images in the SOC dataset. . . . .	52
2.10	Distribution of OCR accuracy values of images in the SOC dataset. . . . .	53
3.1	Overview of the unsupervised FR measure combination method. . . . .	62
3.2	Effect of $\lambda_0$ on SROCC (Tested on LIVE). . . . .	77
3.3	Effect of $\lambda_0$ on LCC (Tested on LIVE). . . . .	77
3.4	Effect of dataset size on SROCC (Tested on LIVE, $\lambda_0 = 4$ ). . . . .	78
3.5	Effect of dataset size on LCC (Tested on LIVE, $\lambda_0 = 4$ ). . . . .	78
3.6	TID2008 test: (a) MOS vs. SSIM (b) MOS vs. SSIM-SS (Synthetic score with SSIM as base measure.)(c) SSIM vs SSIM-SS. . . . .	81

4.1	Contour plot of $I(\mathcal{E}_{ij}, Pr(s_i, s_j))$ as function of the expectation and the standard deviation of $s_i - s_j$ . . . . .	97
4.2	MOS test in the Crowdsourcing experiment. . . . .	100
4.3	PC test in the Crowdsourcing experiment. . . . .	101
4.4	Example of Image Pairs. . . . .	102
4.5	Kendall's $\tau$ in the Crowdsourcing experiment. . . . .	106
4.6	LCC in the Crowdsourcing experiment. . . . .	106
4.7	Number of MOS and PC tests sampled in each iteration. . . . .	107
4.8	Standard deviation of Kendall's $\tau$ . . . . .	107

## List of Abbreviations

CORNIA	Codebook Representation for No-Reference Image Assessment
FR-IQA	Full-Reference Image Quality Assessment
IQA	Image Quality Assessment
LCC	Linear Correlation Coefficient
NR-IQA/NRIQA	No-Reference Image Quality Assessment
OA	opinion-aware
OF	opinion-free
RR-IQA	Reduced-Reference Image Quality Assessment
SROCC	Spearman Rank Order Correlation
SSIM	Structural Similarity Index Measure

## Chapter 1: Introduction

With the tremendous growth in the use of digital images for representing and communicating information, it is important to have quality control systems that can monitor, maintain and enhance image quality [1].

Digital images may be degraded at various stages of their life cycle. Fig. 1.1 shows a document image generation process [2] and degradations may be introduced at each step of this process, including (1) **Creation**: Documents are used to convey information. The creation of a document is a process in which information in the form of symbols is written or printed upon a medium such as paper or palm leaf. Degradations at the creation stage are introduced because of the document medium (e.g. paper translucency/texture), the devices used to create the document (e.g. inadequate/heavy printing and noise in electronic components) and the production process (e.g. typesetting/handwriting imperfections). (2) **Physical degradation**: Once a document is created, it may be subject to various external degradations or manipulations by a human or the environment. These pre-digitization noises are referred to as physical noises in [3], where they are defined as whatever damage the physical integrity and readability of the original information of a document. (3) **Digitization**: Document digitization is the process of generating digital represen-

tations, usually as a discrete set of pixels. A variety of devices can be used for digitization including for example, scanners, mobile phones and cameras. Digitization operations and hardware defects such as paper positioning variations (e.g. skew), pixel sensor sensitivity variations, vibration and other non-uniform equipment motion may further degrade the document image. (4) **Processing:** Processing refers to all types of processing applied to the digital document image after its creation. For example, given a gray scale document image, binarization is often a first step since many document analysis algorithms require a binarized image. The binarization process may introduce binarization noise. For the purpose of efficient storage, lossy compression algorithms such as JPEG or JPEG2k may be applied to document images and introduce compression noise. Furthermore, the quality of transmission network may affect the quality of document image at the receiver side. To recover information from the degradations arising from previous stages, various restoration and enhancement algorithms may be applied. However, an uninformed application of enhancement techniques may further degrade document images. Examples of document image degradations (or degradation sources) are shown in Table 1.1.

Fig. 1.2 shows the typical process of how a natural scene image is generated. Various degradations may be introduced in this process. For example, motion blur may be introduced by the move of image acquisition devices. The sensor pattern noise may be brought in by the imaging sensor. The image compression process may bring in compression noise, such as the “blockiness” and “ringing” artifacts due to JPEG or JPEG2k compression and the transmission system may introduce transmission noise such as package loss degradation, fast-fading degradation (Table

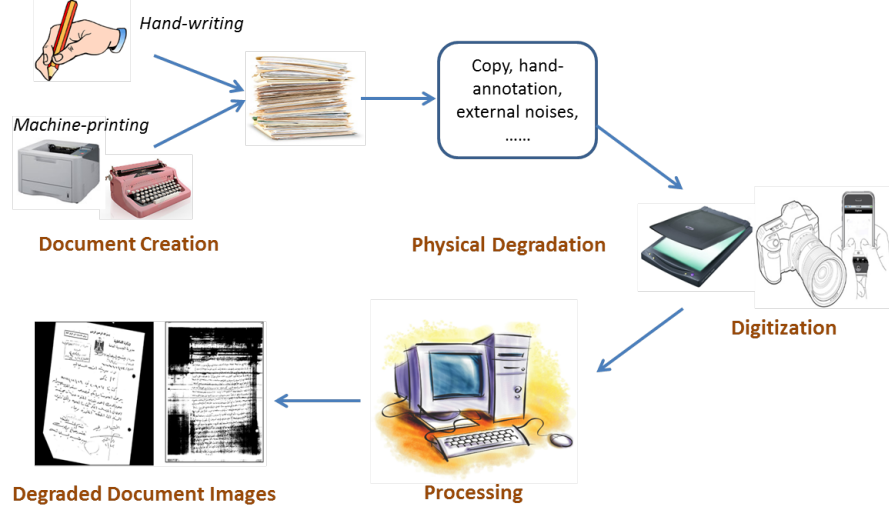


Figure 1.1: Document image generation process.

1.1).

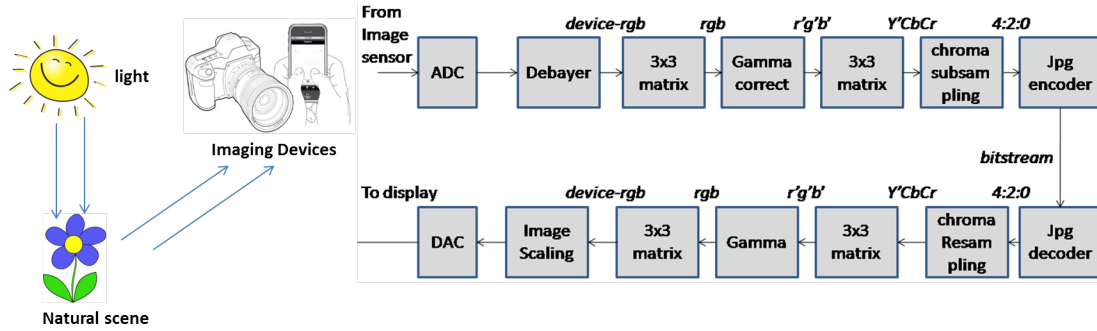


Figure 1.2: Natural scene image generation process.

Figs. 1.3 and 1.4 show examples of distorted natural scene images and document images with various types of distortions. These degradations may lead to the loss of visual information, the poor experience for human viewers and difficulties for image processing and analysis at subsequent stages.

The problem of Image Quality Assessment (IQA) arises in various different image processing and computer vision applications. For example, image process-

	Degradations
Document	Paper translucency; Paper texture; Inadequate printing; Heavy printing; Non-uniform illumination; Low print contrast; Typesetting imperfections; Touching character in handwritten document; Vibration and other non-uniform equipment motion; Noise in electronic components; Bleed-through, shadow-through; Folding marks; Paper aging; Paper Punching; Stains; Thornoff regions; Worm holes; Readers annotations; Carbon copy effect; Scratches and cracks; Sunburn; Defocusing; Paper positioning variations (skew, translation, etc.); Pixel sensor sensitivity variations; Vibration and other non-uniform equipment motion; Noise in electronic components; Irregular pixel sensor placement (e.g. not lying on a perfectly square grid); Finite spatial sampling rate; Non-flat paper surface (e.g. curling and warping); Non-rectilinear camera positioning (e.g. perspective distortion); Binarization; Document enhancing; JPEG/JPEG2K compression; transmission noise, etc.
Natural Scene	Gaussian Blurring; White Gaussian Noise; JPEG Compression; JPEG2K Compression; FastFading distortion, JPEG transmission errors; JPEG2K transmission errors; Impulse Noise; Quantization Noise; Salt-and-Pepper Noise, etc.

Table 1.1: Examples of image degradations or image degradation sources.

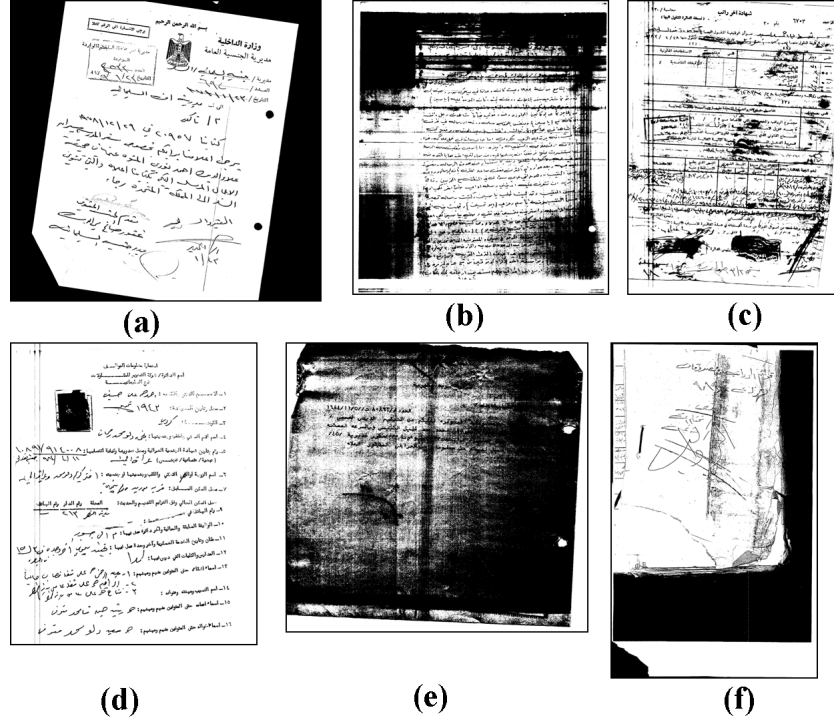


Figure 1.3: Examples of distorted document images.



Figure 1.4: Examples of distorted natural scene images (a) Undistorted reference image. (b) JPEG2000 Compression. (c) JPEG Compression. (d) White Gaussian Noise. (e) Gaussian Blur. (f) Fast Fading.



ing and transmission systems may have their parameters be adjusted according to image quality measures; image retrieval systems can use quality measures to rank images and image processing algorithms may be evaluated by a quality measure; a document image processing system may use quality measure to automatically filter out pages with low predicted OCR accuracy and guide the selection of document image enhancement methods.

There are two divisions of IQA research: objective IQA and subjective IQA. The goal of objective IQA is to develop a computational model that can predict the quality of distorted image with respect to human perception or other measures of interest accurately and automatically. Based on the availability of reference images, objective IQA approaches can be broadly classified into: full-reference (FR), no-reference (NR) and reduced-reference (RR) approaches. When the reference images are available, FR approaches can be applied to directly quantify the differences between distorted images and their undistorted ideal versions. State-of-the-art FR measures yield high correlation with human perception, however, FR approaches cannot be used in many practical applications where there do not exist such reference images. To address this problem, there has been an increasing interest in developing NR approaches, which do not require any information of the reference image to compute the quality measure. NR approaches can be further classified into two categories: distortion-specific (DS) approaches and general-purpose approaches. DS approaches for NR-IQA usually target one or two specific types of distortions and prior knowledge on the distortion properties is embedded in algorithm designs. Unlike DS approaches, general-purpose NR-IQA approaches do not investigate any

particular type of distortion but rather build a general computational model to work universally for different types of distortions. In addition to FR and NR approaches, RR approaches which lie between the FR and the NR approaches have also been extensively studied. RR approaches do not require the full information of the reference image, but partial information extracted from the reference image is used to quantify the image degradations. It is useful in a number of applications, for example, in real-time visual communication systems, we may use RR approaches to track image quality degradations throughout the communication process and adjust the system parameters or allocate resources according to image quality. A summary of different types of objective IQA approaches is shown in Fig. 1.5, where tasks studied in this thesis are highlighted.

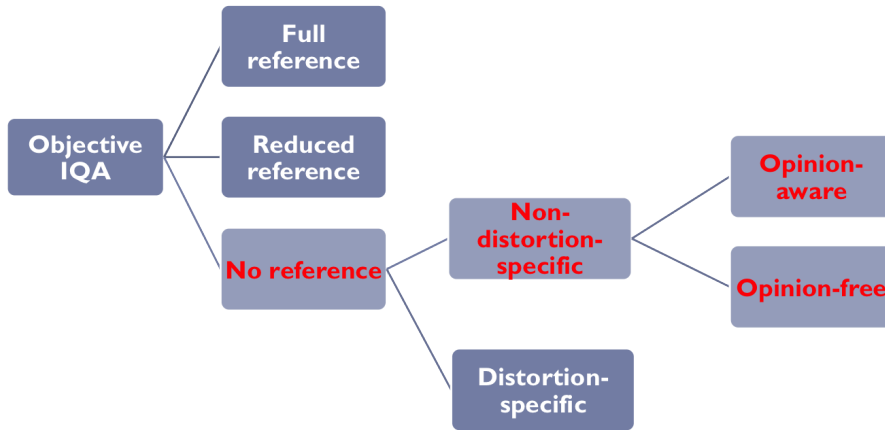


Figure 1.5: Objective IQA categories.

The second main division of IQA research is subjective IQA. Subjective IQA tests ask humans to evaluate the quality of images and are generally considered as the most reliable way to evaluate the visual quality of digital images perceived by

the end user. It is often used to construct image quality datasets and provide the groundtruth for building and evaluating objective quality measures. Two types of subjectives tests have been introduced in previous studies: the Absolute Categorical Rating (ACR) test and the Paired Comparison (PC) test. Both types of tests have their limitations and advantages. How to efficiently and effectively conduct subjective IQA test is still an open problem. Conventional subjective IQA experiments conducted in laboratory settings are usually expensive and time-consuming and typically only a small number of subjects are involved. With the ubiquitous internet access and the rise of internet micro-labor markets such as Amazon Mechanical Turk, there has been an increasing interest in designing subjective IQA tests for crowdsourced settings. The use of crowdsourcing for subjective IQA brings in new opportunities and challenges for this problem.

In this thesis, we will address the following three problems of IQA:

1. How can we automatically learn discriminative features for NR-IQA tasks?
2. How can we train a NR-IQA model without using human opinion scores?
3. How can we design the subjective test so that we can achieve desired accuracy with minimal cost?

## 1.1 Feature Learning for No-Reference Image Quality Assessment

We first study objective IQA and focus on the most challenging objective IQA task: general-purpose No-Reference (NR) IQA. The objective of general-purpose NR-IQA is to build a computational model which can automatically predict hu-

man (or machine) perceived quality of digital images without using the undistorted reference image and without examining the properties of specific distortions. Additionally, our goal is to develop a general-purpose NR-IQA model that can be used in real-time applications and across different image domains.

Previous approaches to this problem typically rely on hand-crafted features which are carefully designed based on prior knowledge. Tables 1.2 and 1.3 summarizes typical features that have been used in previous works for natural scene IQA and document IQA.

Table 1.2: Features for objective Natural Scene IQA.

Feature	Reference
Phase Congruency	[4]
Image Entropy	[4]
Image Gradient	[4]
NSS Feature in complex pyramid wavelet transform domain	[5]
Cross-scale distribution of wavelet coefficient phase	[5]
Patch PCA singularity	[5]
Two-color prior based blur statistics	[5]
Wavelet domain NSS Feature	[6]
DCT domain NSS Feature	[7]
Spatial domain NSS Feature	[8]
Mean, variance and entropy of wavelet coefficients	[9]

In contrast, we use raw-image-patches extracted from a set of unlabeled images to learn a dictionary in an unsupervised manner [19]. We use soft-assignment coding with max pooling to obtain effective image representations for quality estimation. Our algorithm is computationally appealing, using raw image patches as local descriptors and using soft-assignment for encoding. Furthermore, unlike previous methods, our unsupervised feature learning strategy enables our method

Table 1.3: Features for objective Document IQA.

Feature	Reference
Font Size	[10]
Normalized Font Size	[11]
Stroke Thickness	[10, 12]
Small Speckle Factor	[10, 11]
White Speckle Factor	[10, 11, 13]
Touching Character Factor	[10, 11]
Broken Character Factor	[10, 11, 13]
Number of pixels	[14]
Average width of CC	[14]
Average height of CC	[14]
Number of CC	[14]
Pixel density	[14]
Gradient of the edge	[15]
Average height-width ratio	[15]
Foreground/Background Uniformity	[12]
Sharpness	[12]
Transient Region Density	[12]
Stability of CC values	[12]
Continuity	[12]
Noise measure based on Median Filtering	[12]
Pulse width ratio	[12]
Entropy	[12]
Stroke density distribution	[16]
Histogram	[16]
Crossing Count	[16]
Morphological-based features	[17]
Noise-removal-based features	[17]
Spatial characteristics features	[17]
$\Delta DoM$	[18]

to adapt to different domains. Our system, CORNIA (Codebook Representation for No-Reference Image Assessment), has been tested on the LIVE database and performs statistically better than the full-reference quality measures Peak-Signal-to-Noise-Ratio (PSNR) and structural similarity index (SSIM) and is shown to be comparable to state-of-the-art general purpose NR-IQA algorithms.

To boost the performance of CORNIA, we further introduce a supervised feature learning method for CORNIA [20]. Previous approaches including the unsupervised CORNIA treat local feature extraction and regression model training independently, but supervised CORNIA utilizes back-projection to link the two steps and learns a compact set of filters which can be applied to local image patches to obtain discriminative local features. Using a small set of filters, supervised CORNIA is extremely fast. Overviews of different NR-IQA systems are presented in Fig. 1.6 to demonstrate the differences between conventional NR-IQA approaches, the unsupervised CORNIA and the supervised CORNIA.

## 1.2 Blind Learning of Image Quality based on Synthetic Scores

State-of-the-art general purpose NR-IQA methods rely on 1) examples of distorted images and 2) corresponding human opinion scores to learn a regression function that maps image features to the quality score. These types of models are considered “opinion-aware” (OA) NRIQA models. A large set of human scored training examples is usually required to train a reliable OA-NRIQA model. However, obtaining human opinion score through subjective testing is often expensive and time-consuming. It is therefore desirable to develop “opinion-free” (OF) NR-IQA models that do not require human opinion scores for training.

To approach this challenge, we introduce BLISS (Blind Learning of Image Quality using Synthetic Scores) [21]. BLISS is a simple, yet effective method for extending OA-NRIQA models to OF-NRIQA models. Instead of training on human

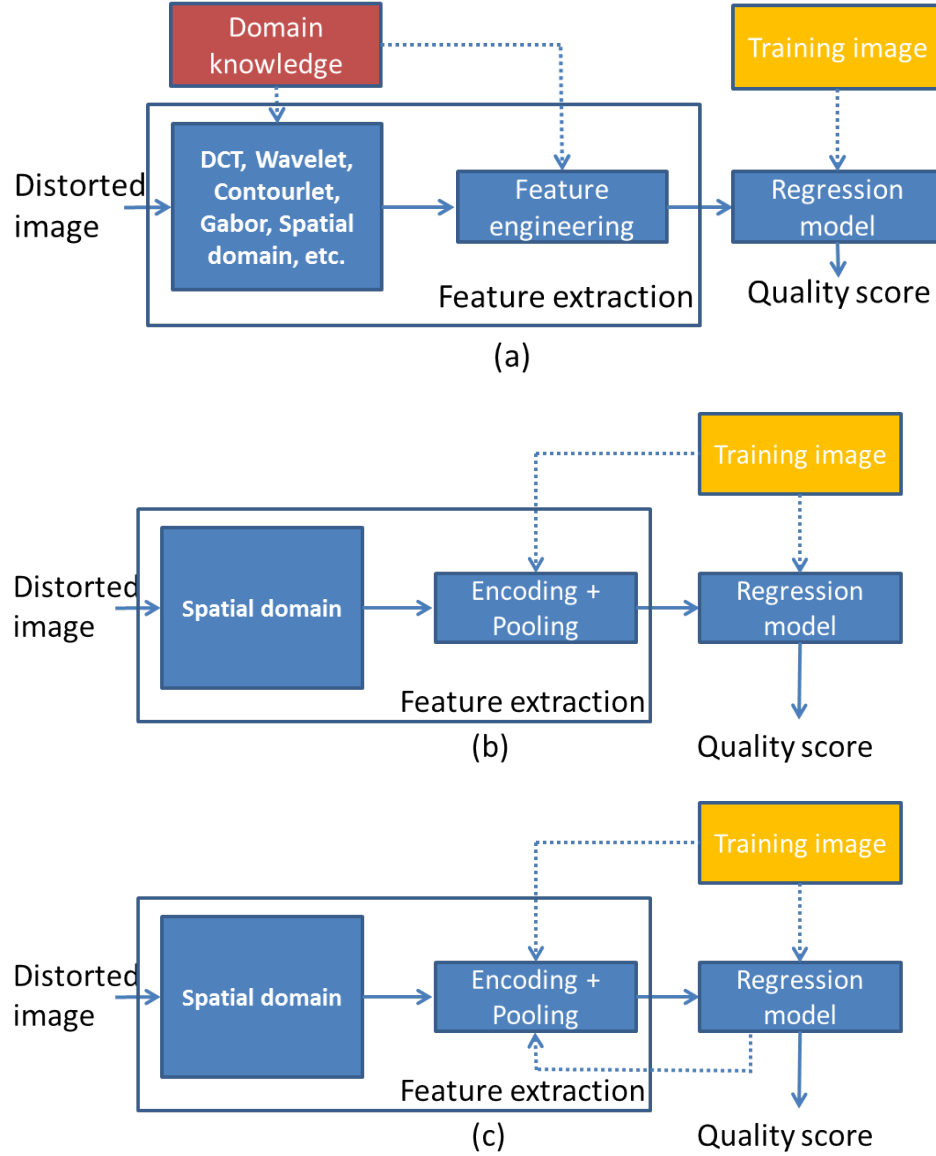


Figure 1.6: Overview of NR-IQA systems (a) Conventional NR-IQA systems. (b) Unsupervised CORNIA. (c) Supervised CORNIA.

opinion scores, we train NR-IQA models on Full-Reference (FR) IQA measures. State-of-the-art FR measures yield high correlation with human opinion scores, and can therefore serve as an approximation to human opinion scores. Unsupervised rank aggregation is applied to combine different FR measures to generate a synthetic score, which serves as a better “gold standard”. Extensive experiments on three

standard IQA datasets show that BLISS significantly outperforms previous OF-NRIQA methods and is comparable to state-of-the-art OA-NRIQA methods.

### 1.3 Active Learning for Subjective Image Quality Assessment

Subjective Image Quality Assessment (IQA) is the most reliable way to evaluate the visual quality of digital images perceived by the end user. It is often used to construct image quality datasets and provide the groundtruth for building and evaluating objective quality measures. Subjective tests based on the Absolute Categorical Rating (ACR), for example the Mean Opinion Score (MOS) test, have been widely used in previous studies, but have many known problems such as an ambiguous scale definition and dissimilar interpretations of the scale among subjects. To overcome these limitations, Paired Comparison (PC) tests have been proposed as an alternative and are expected to yield more reliable results. However, PC tests can be expensive and time consuming, since for  $n$  images they require  $\binom{n}{2}$  comparisons.

We present a hybrid subjective test which combines MOS and PC tests via a unified probabilistic model and an active sampling method [22]. Our method actively constructs a set of queries consisting of MOS and PC tests based on the expected information gain provided by each test and can effectively reduce the number of tests required for achieving a target accuracy. Our method can be used in conventional laboratory studies as well as crowdsourcing experiments. Experimental results show our method outperforms state-of-the-art subjective IQA tests in a crowdsourced setting.



## 1.4 Outline

The remainder of this dissertation is organized as follows. Chapter 2 introduces unsupervised and supervised CORNIA – a feature learning framework for NR-IQA. Chapter 3 describes BLISS – an effective method for extending “opinion-aware” NR-IQA models to “opinion-free” NR-IQA modes. In Chapter 4, we present a hybrid subjective IQA test which combines the MOS and the PC tests via active sampling. Chapter 5 summarizes our work.

## Chapter 2: Feature Learning for No-Reference Image Quality Assessment

### 2.1 Introduction

This chapter addresses the problem of general-purpose No-Reference Image Quality Assessment (NR-IQA) with the goal of developing a real-time, cross-domain model that can predict the quality of distorted images without prior knowledge of non-distorted reference images or the types of distortions present in these images. NR-IQA has long been considered as one of the most difficult problems in image analysis [23]. Recently, however, significant progress has been made in the field. State-of-the-art general-purpose NR-IQA systems [5–8, 19, 20] have been shown to outperform FR measures: Peak Signal-to-Noise ratio (PSNR) and Structural Similarity Index Measure (SSIM) on standard IQA datasets. Previous approaches however have a number of shortcomings.

First, existing state-of-the-art algorithms typically rely on hand-crafted features which are carefully designed based on prior knowledge. The use of hand-crafted features may limit the applications of these approaches in practice.

Second, speed is an important issue for NR-IQA systems since NR-IQA mea-

asures are often used in real-time imaging or communication systems. For example, IQA measures may be embedded in visual communication systems to help optimize the system parameters and guide the allocation of network resources in real time. Algorithms that rely on computationally expensive image transforms [5–7] cannot be used in these applications. Therefore, it is desirable to develop a NR-IQA system that can be used in real-time systems.

Third, previous work on NR-IQA has focused primarily on natural scene images and image quality is defined with respect to human perception. Very limited work has been done for NR-IQA for other types of images, such as camera-captured or scanned document images. Document IQA has been found to be very useful in many document image processing applications. For example, depending on the level of degradation, the performance of modern OCR systems may suffer. Document IQA can help to automatically filter pages with low predicted OCR accuracy or guide the selection of document image enhancement methods. Conventional image quality measures developed for natural scene images do not work well for document images since document images have very different characteristics than natural scene images. For example, most document images are binary and gray-scale consisting of black text and white background. Building a NR-IQA system that can be adapted to images with different characteristics is a challenging problem.

To overcome these limitations, the objective of this work is: first, to develop a fast NR-IQA method that can be used in real-time systems and second, to develop a general learning-based framework that can be applied to various different image domains. To approach these challenges, we developed a unified NR-IQA frame-

work called CORNIA (COdebook Representation for No-reference Image quality Assessment). CORNIA is a feature learning framework for NR-IQA, in which discriminative features are directly learned from raw image pixels. Both unsupervised and supervised feature learning approaches have been studied under our framework.

### 2.1.1 Related work

#### **Natural Scene Statistics for NR-IQA**

Natural Scene Statistics (NSS) based approaches have been successfully applied to IQA for natural scene images. These methods are based on the following observations. First, when images are properly normalized or transferred to some transform domains (e.g. DCT or wavelet domain), local descriptors (e.g. normalized intensity values, wavelet coefficients, etc.) can be modeled by some parametric distributions. Second, the shape of these distributions are very different for non-distorted and distorted images. These fundamental observations form basis of many recent IQA approaches [6–8], which differ from each other primarily in how the local descriptors are extracted. For example, DIIVINE [6] extracts local descriptors in wavelet domain. Cosine transform coefficients based descriptors are used in BLIINDS-II [7]. BRISQUE [8] directly models the normalized image pixel value using generalized Gaussian distributions (GGD) and models product of neighboring pixels by asymmetric generalized Gaussian distributions (AGGD).

The success of these methods rely largely on how local features are computed, therefore hand-crafted features designed specifically for a particular domain are often

used. This limits their applications in other image domains.

### **Unsupervised Feature Learning**

With the increasing availability of computational resources, there has been a greater emphasis on unsupervised feature learning. The goal of unsupervised feature learning is to automatically learn a good representation of the input from unlabeled data instead of hand-engineering feature representations. Most previous work has focused on applying unsupervised feature learning to classification problem. We apply it to NR-IQA, a regression problem. It serves as a case study for applying unsupervised feature learning to regression problems in general.

### **Supervised Feature Learning**

As a natural extension to the unsupervised feature learning based method, we have also developed a supervised feature learning method for NR-IQA. The supervised filter learning method is closely related to supervised dictionary learning for image classification. Earlier methods for dictionary learning focused on reconstruction of signals and ignored label information. To learn a more compact and discriminative dictionary, learning approaches that jointly optimize both a reconstructive and a discriminative criterion have been developed [24–26]. Unlike conventional supervised dictionary learning, which requires the linear combination of the learned atoms in dictionary be able to well represent image patches, we do not have this constraint in our supervised filter learning process. In fact, it will be shown later that the functionality of filter for NR-IQA and codeword for image classification are very different.

Supervised filter learning has also been explored by Jain and Karu in [27] for

texture classification, where feature extraction and classification tasks are performed by a neural network. The learned filters are weight vectors in the first layer of the network. Our supervised feature learning method also learns a compact set of filters using a back-propagation approach but differs in the final stage where we perform support vector regression (SVR) using learned filters for predicting image quality.

### 2.1.2 Our approach

CORNIA advances previous approaches in two important ways. First, we use raw image pixels based local descriptors in our learning framework, which are efficient and easily computable. Using supervised CORNIA, real-time computation can be achieved. In contrast, previous state-of-the-art general purpose NR-IQA algorithms [6, 7, 28, 29] use off-the-shelf image transformation and filtering techniques such as wavelet transform, cosine transform and Gabor filtering for extracting features, which can be very time consuming. Second, if the domain of the problem changes, say from natural scene images to document images, the performance of previous techniques is unpredictable, while our method is based on feature learning and does not embed any prior knowledge of the domain, making it more general and giving it the potential to adapt to different domains. An overview of differences between our method and previous methods is illustrated in Fig. 1.6.

## 2.2 Feature Learning for NR-IQA

The feature learning framework adopted in this work is illustrated in Fig. 2.1. It consists of the following components (1) a local feature extractor; (2) a global feature extractor, which summarizes the distribution of local features and (3) a regression model. Both unsupervised and supervised feature learning methods have been studied in this framework.

In the unsupervised feature learning system, local descriptors are encoded using filters that are learned in an unsupervised way from a set of unlabeled distorted images. By using a large set of filters (usually on order of thousands), the unsupervised method can capture different aspects of distortions and accurately predict the quality of distorted image. However, when only a small set of filters is used, the performance of this method may drop significantly. To improve the speed and reduce the redundancy in the learned feature representations, we further develop a supervised feature learning method, in which a compact set of filters is learned by jointly optimizing the local feature extractor and the regression model in a supervised way. The learned compact set of filters yields more discriminative features when applied to local image patches. Using a small set of filters, the feature extraction process is much faster and more memory efficient.

Inputs to our system are unlabeled images for learning an initial set of filters and labeled images for learning the prediction model (in our case, a linear Support Vector Regression (SVR) model). The system outputs the trained regression model and a set of filters which are used for extracting discriminant local features.

When unsupervised feature learning is used, the dashed arrow ( Fig. 2.1) which links the regression model and the encoding module is discarded, and the output filters are learned without feedback from the regression model. In the supervised feature learning method, prediction errors obtained from the regression model are sent back to the encoding module for filter updating and the output filters are learned in a supervised way.

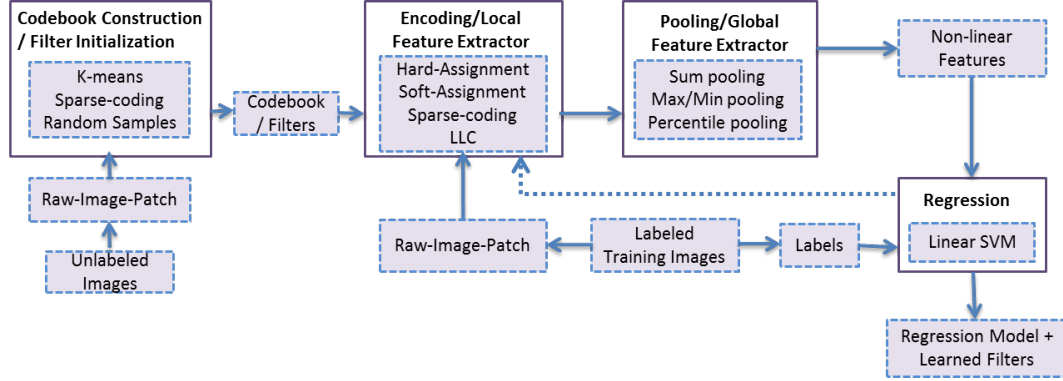


Figure 2.1: Learning Features and Prediction Model for NR-IQA.

As shown in Fig. 2.1, we can choose different implementations for each module in the system. For example, the initial filters can be learned using k-means or sparse coding and we can use hard encoding, “Localized” soft-assignment coding [30], sparse coding [31] or Locality-constrained Linear Coding (LLC) [32] in the encoding module. Although different choices can be explored, in this thesis, we describe only specific implementations which we have found to be both efficient and effective. In the remainder of this section, we describe the feature extraction process with fixed a set of filters, discuss the unsupervised feature learning method and finally describe our supervised feature learning method.



### 2.2.1 Feature extraction

In this section, we discuss how a set of linear filters is used to obtain global features. Suppose an image is represented by a set of local descriptors, where these local descriptors are normalized raw image patches:

$$X = [x_1, x_2, \dots, x_N] \in R^{d \times N}$$

where the column vector  $x_i$  denotes the  $i$ -th local descriptor of the image. The normalization is performed by subtracting the local mean value from each patch, and dividing it by its standard deviation. This normalization process is similar to that used in [8] and it is worth noting that it is essential for performance and we do not consider contrast changes and intensity shifts as degradations.

A set of filters is represented by  $B = [b_1, \dots, b_K] \in R^{d \times K}$ , where the column vector  $b_i$  ( $\|b_i\|_2^2 = 1$ ) denotes the  $i$ -th filter and  $K$  is the number of filters.

#### 2.2.1.1 Local feature encoding

The first step in this work-flow is local feature encoding using linear filters. Specifically, each local descriptor is encoded by its responses to the set of linear filters (i.e. inner product between local descriptors and filters). Details of how to learn these linear filters will be described later.

Assume we have an image level representation matrix as follows:

$$\Omega = B^T \times X = \begin{pmatrix} b_1 \cdot x_1 & b_1 \cdot x_2 & \cdots & b_1 \cdot x_N \\ b_2 \cdot x_1 & b_2 \cdot x_2 & \cdots & b_2 \cdot x_N \\ \vdots & \vdots & \ddots & \vdots \\ b_K \cdot x_1 & b_K \cdot x_2 & \cdots & b_K \cdot x_N \end{pmatrix} \quad (2.1)$$

Examples of distributions of filter responses from images with different types and levels of distortions are shown in Fig. 2.2 (with the reference image shown in Fig. 1.4). We can see from this figure that with properly learned filters, statistics extracted from these distributions can be good indicators of image quality. The *filter* here is similar to the *codeword* in the image classification literature in that both are used for local feature encoding. However, their functionalities are very different. For example, in the object recognition problem, a *codeword* resembles a part of the object, and the maximal response to each codeword indicates its presence or absence in the image. In our problem, both maximal and minimal responses are informative and important for the prediction task. More generally, we are interested in characterizing the entire distribution of the filter responses.

Our encoding method can be considered a type of soft-assignment encoding. We may use other encoding methods in this step. However, we will show experimentally that the simple soft-assignment encoding is comparable to more complicated encoding methods [30–32] and is much faster to compute.

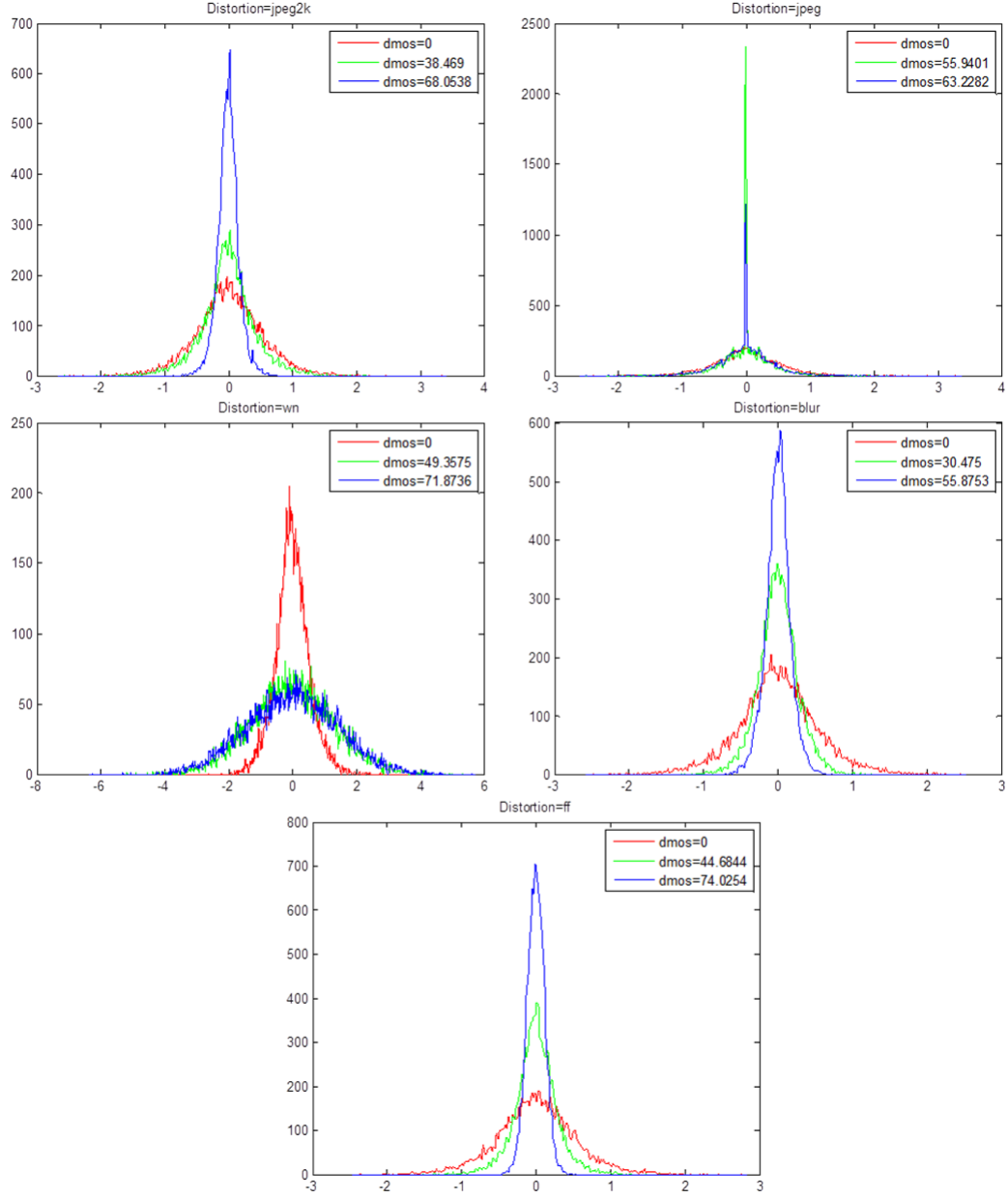


Figure 2.2: Examples of filter responses for different types and levels of distortions (High DMOS indicates low quality).

### 2.2.1.2 Summarizing statistics

Statistics which summarize the distribution of local features are extracted as global descriptors. Specifically, we use the maximal and minimal values of filter

responses for describing the effect of filters on the image. The image level descriptor of  $X$  can be written as:

$$Z = [\max(\Omega)^T, \min(\Omega)^T] \quad (2.2)$$

where  $\max$  and  $\min$  operations are applied on each row of  $\Omega$  and superscript  $T$  means transpose. Other statistics which summarize the distribution of filter responses can also be explored, such as skewness and kurtosis. One can also use a parametric model, for example, BRISQUE [8] models normalized intensity values using GGD and the shape and scale parameters are used as features. In our framework, the supervised filter learning process needs to compute the gradient of the summarizing statistics with respect to the encoding filter. We therefore require statistics with simple analytical forms or at least statistics that have good analytical approximations. Although the minimal and maximal value of filter responses may not be accurate in characterizing the shape of distribution for discriminating images with high and low quality, they are fairly good indicators (Fig. 2.2), in addition to being efficient to compute.

Combining the two steps above, we have  $Z = \phi(X, B)$  where  $Z \in R^{2K \times 1}$  with the first  $K$  elements corresponding to maximal responses and the last  $K$  elements corresponding to minimal responses. The step of extracting summarizing statistics is also known as visual feature pooling, which is a research topic that has been extensively studied in the image classification literature.

### 2.2.2 Unsupervised Filter learning

We have introduced how to use a set of linear filters to extract local features. Next, we introduce how to learn a set of linear filters such that the encoded local features yield discriminant and effective global features for the NR-IQA task. We first introduce an unsupervised method for filter learning. A supervised extension will be described in Section 2.2.3.

Following the convention in [19], we use the term *codebook* to refer to a set of filters. In unsupervised CORNIA, the visual codebook is constructed by performing K-means clustering on local features extracted from unlabeled training images. A matrix  $B_{d \times K} = [b_1, b_2, \dots, b_K]$  denotes a visual codebook, where  $b_{i(i=1, \dots, K)}$  are centroids of clusters learned by K-means clustering. More complex training methods such as sparse coding (SC) can be used to perform codebook construction (or dictionary learning) to improve system performance, but the use of K-means clustering in our work is motivated by the observation in [33] that a good encoding scheme is more critical than dictionary learning.

The learned codebook is normalized so that each of these bases has unit length. Examples of learned filters are shown in Fig. 2.3. Filters with the “dot” patterns come from patches with salt-pepper noise, while “smooth” filters correspond to blurred patches and filters with horizontal and vertical line patterns correspond to patches with “blockiness”. As shown in Fig. 2.3, some filters learned in this way resemble Gabor filters.

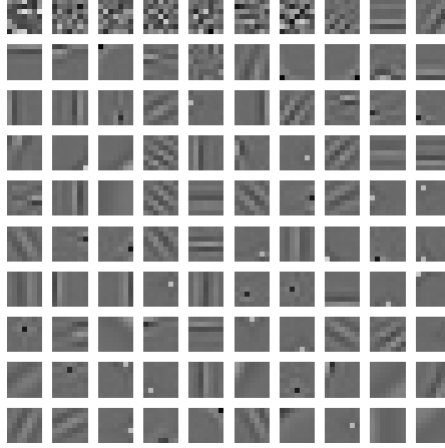


Figure 2.3: Randomly selected centroids trained on CSIQ database using K-means.

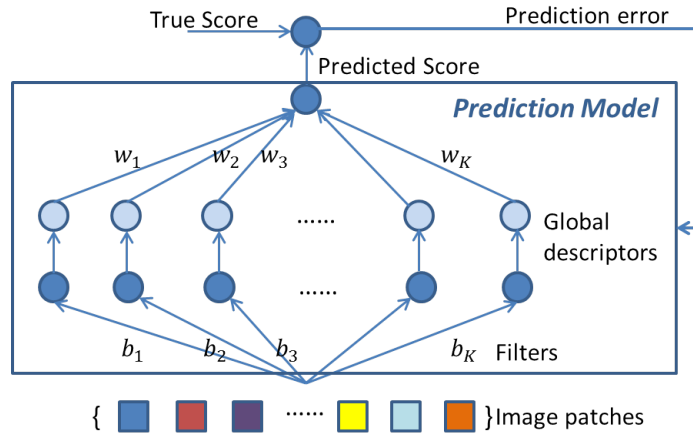


Figure 2.4: Overview of supervised filter learning method with linear SVR.

### 2.2.3 Supervised Filter Learning

We now introduce a supervised filter learning method in which a compact set of filters is learned by jointly optimizing the local feature extractor and the regression model in a supervised way. An overview of the system is shown in Fig. 2.4.

### 2.2.3.1 Problem formulation

Suppose we have  $n$  training images and the  $k$ -th training image is denoted as  $X_k$  with a corresponding feature vector denoted as  $Z_k$ . Its regression target, i.e, the true quality score, is denoted as  $y_k$ . For supervised filter learning, we use linear  $\epsilon$ -Support Vector Machine Regression ( $\epsilon$ -SVR) for training. The prediction function takes the following form:

$$f(Z_k, w) = \sum_{i=1}^{2K} w_i Z_k(i) + w_0$$

where  $Z_k(i)$  is the  $i$ -th element in  $Z_k$  and  $w$  is learned by minimizing the sum of a loss function and a regularization term:

$$\min_w \left\{ \sum_{k=1}^n L(y_k, f(Z_k, w)) + \lambda_1 \|w\|_{l_2}^2 \right\}$$

where  $L$  is the  $\epsilon$ -insensitive loss function described by:

$$L(y, \hat{y}) = \begin{cases} 0 & \text{if } |y - \hat{y}| \leq \epsilon \\ |y - \hat{y}| - \epsilon & \text{otherwise} \end{cases} \quad (2.3)$$

In the above formulation, the prediction model is trained with the set of filters fixed. Our supervised filter learning method jointly optimizes the prediction model and the set of filters. The objective function of this optimization problem is defined

as follows:

$$\begin{aligned}
C(B, w, \{X_k\}_{k=1}^n) &= \sum_{k=1}^n L(y_k, f(\phi(X_k, B), w)) \\
&+ \lambda_1 \|w\|_{l_2}^2 + \lambda_2 \text{avecorr}(B) \\
&\text{subject to } \|b_i\| = 1, i = 1, \dots, K
\end{aligned} \tag{2.4}$$

where  $\lambda_1$  is a balancing factor of the regularization term in the prediction model and  $\text{avecorr}(B) = \frac{1}{K-1} \sum_{i=1}^K \sum_{j:j \neq i} < b_i, b_j >$  is the average correlation of one filter with every other filter. This correlation penalty term is added to avoid learning highly correlated filters. Furthermore, in order to avoid the over-fitting problem and regularize the search space of the optimal filters, we add the constraint that  $\|b_i\| = 1 (i = 1, \dots, K)$ .

Optimal  $B$  and  $w$  is given by

$$(B^*, w^*) = \text{argmin}_{B, w} C(B, w, \{X_k\}_{k=1}^n)$$

This optimization problem can be solved by optimizing alternatively over  $B$  and  $w$ . The initial set of filters are obtained by performing k-means clustering on a set of local features. When  $B$  is fixed, an optimal  $w$  can be found using a standard SVR program [34]. Given  $w$ , we can apply stochastic gradient descent (SGD) to find the optimal  $B$ .

### 2.2.3.2 Optimizing B

#### Computing the Gradient

The SGD process requires to compute the gradient of the objective function  $C$  with



respect to a filter  $b_i, i = 1, \dots, K$ . Using chain rule, we can have:

$$\frac{\partial C}{\partial b_i} = \sum_{k=1}^n \frac{\partial L}{\partial f_k} \frac{\partial f_k}{\partial Z_k} \frac{\partial Z_k}{\partial b_i} + \lambda_2 \frac{\partial \text{avecorr}(B)}{\partial b_i} \quad (2.5)$$

where  $f_k = f(Z_k, w)$  is the predicted quality score for the  $k$ -th training image.

The loss function used in  $\epsilon$ -SVR is non-differentiable, therefore we use the following approximation (Huber loss) for computing the gradient

$$L(y, \hat{y}) = \begin{cases} 0 & |y - \hat{y}| \leq \epsilon - h \\ |y - \hat{y}| - \epsilon & |y - \hat{y}| \geq \epsilon + h \\ \frac{(y - \hat{y} - \epsilon + h)^2}{4h} & \epsilon - h < y - \hat{y} < \epsilon + h \\ \frac{(y - \hat{y} + \epsilon - h)^2}{4h} & -\epsilon - h < y - \hat{y} < -\epsilon + h \end{cases} \quad (2.6)$$

where  $0 < h < \epsilon$ . When  $h \rightarrow 0$ , Eq. 2.6 is equivalent to the  $\epsilon$ -insensitive loss used in  $\epsilon$ -SVR. The derivative of the above loss function is given by:

$$\frac{\partial L}{\partial f_k} = \begin{cases} 0 & |y_k - f_k| \leq \epsilon - h \\ -1 & y_k - f_k \geq \epsilon + h \\ 1 & y_k - f_k \leq -\epsilon - h \\ \frac{f_k - y_k + \epsilon - h}{2h} & \epsilon - h < y_k - f_k < \epsilon + h \\ \frac{f_k - y_k - \epsilon + h}{2h} & -\epsilon - h < y_k - f_k < -\epsilon + h \end{cases} \quad (2.7)$$

The derivative of the prediction function with respect to the global feature

vector is given by <sup>1</sup>:

$$\frac{\partial f_k}{\partial Z_k} = [w_1, w_2, \dots, w_{2K}] \quad (2.8)$$

The global feature vector  $Z_k = [Z_k(1), \dots, Z_k(2K)]^T$ , where for  $i = 1, \dots, K$  is:

$$Z_k(i) = b_i \cdot x_{max,i}^k, \quad x_{max,i}^k = \operatorname{argmax}_{x_l \in X_k} (b_i \cdot x_l)$$

$$Z_k(i+K) = b_i \cdot x_{min,i}^k, \quad x_{min,i}^k = \operatorname{argmin}_{x_l \in X_k} (b_i \cdot x_l)$$

where superscript  $k$  is the index of the training image,  $x_l \in X_k$  means  $x_l$  is a local feature vector from image  $X_k$  and  $\cdot$  represents inner product. We therefore have  $\frac{\partial Z_k(i)}{\partial b_i} = x_{max,i}^k$ ,  $\frac{\partial Z_k(i+K)}{\partial b_i} = x_{min,i}^k$  and the derivative of the global feature vector with respect to  $b_i$  is given by:

$$\frac{\partial Z_k}{\partial b_i} = [0, \dots, 0, \frac{\partial Z_k(i)}{\partial b_i}, 0, \dots, 0, \frac{\partial Z_k(i+K)}{\partial b_i}, 0, \dots, 0]^T \quad (2.9)$$

$$= [0, \dots, 0, x_{max,i}^k, 0, \dots, 0, x_{min,i}^k, 0, \dots, 0]^T$$

The derivative of the correlation penalty term with respect to  $b_i$  is given by:

$$\frac{\partial \operatorname{avecorr}(B)}{\partial b_i} = \frac{1}{K-1} \sum_{j:j \neq i} b_j^T \quad (2.10)$$

In summary, when linear  $\epsilon$ -SVR is used, we can compute the derivative of the

---

<sup>1</sup>If  $y \in R^m$ ,  $x \in R^n$ , then  $\frac{\partial y}{\partial x} \in R^{m \times n}$

objective function as follows:

$$\frac{\partial C}{\partial b_i} = \left( \sum_{k=1}^n \frac{\partial L}{\partial f_k} (w_i x_{max,i}^k + w_{i+K} x_{min,i}^k) + \lambda_2 \frac{1}{K-1} \sum_{j:j \neq i} b_j \right)^T \quad (2.11)$$

### Stochastic Gradient Descent

Our optimization problem has the constraint that  $\|b_i\|_{l_2}^2 = 1$ , so we perform SGD on the unit sphere. This can be done by projecting the gradient on the tangent plane of the sphere. We describe the SGD process for optimizing one filter as follows:

1. Permute the training images randomly, set  $k = 1$  and initialize  $b^1$ .
2. Compute the gradient  $g_k = \nabla_b C^k(b)|_{b=b^k}$ , where  $C^k = L(y_k, f(Z_k, w)) + \lambda_2 \text{avecorr}(b^k)$  and  $b^k$  is the value of the filter at the  $k$ -th iteration.
3. Project  $g_k$  on the tangent plane of the unit sphere at  $b^k$ ,  $h_k = g_k - (g_k \cdot b^k)b^k$  and normalize it,  $n_k = h_k/|h_k|$ .
4. Update  $b^k$  with  $b^{k+1} = b^k \cos(r_k) + n_k \sin(r_k)$ , where  $r_k$  is the learning rate at the  $k$ -th iteration and  $k = k + 1$ .
5. Go to step 2 and repeat the process until the maximal number of iterations is reached.

### Early Termination

The back projection based method may suffer from the over-fitting problem. In order to avoid over-fitting, we adopt the early termination criteria, proposed in [35]. Specifically, we divide the training data into a training set and a validation set. In each iteration of the optimization process, we train on the training set and evaluate results on the validation set. The training process is terminated as soon as the error

on validation set satisfies our early stopping rule. The set of filters which gives the best performance on the validation set is chosen as the output of the optimization process. For NR-IQA problems, the linear correlation coefficient (LCC) is typically used as an evaluation measure. So we define the following generalization loss based on the LCC. The generalization loss at the  $k$ -th iteration is given by

$$GL(k) = 100((1 - corr(k))/(1 - corr_{opt}(k)) - 1)$$

where  $corr(k)$  is the LCC on validation set at the  $k$ -th iteration and  $corr_{opt}(k) = \max_{k' \leq k} corr(k')$ . Two stopping rules are used:

**(Rule 1)**  $GL(k) > \alpha$  for some  $\alpha > 0$ ;

**(Rule 2)**  $corr$  decreases in consecutive  $l$  iterations.

Training terminates if Rule 1 or Rule 2 is true.

## 2.3 Experiments

In this section, we present experimental results on both natural scene images and document images.

### 2.3.1 Protocol

First, we describe the protocol used in our experiments. Notationally,  $CBk$  will be used to refer the unsupervised CORNIA with  $k$  filters and  $SFk$  for the supervised CORNIA with  $k$  filters. For example  $CB100$  refers to unsupervised CORNIA with 100 filters.

## Experimental settings

In our experiments, we fix parameters for the filter learning, patch extraction and cross-validation as follows:

For supervised filter training, on the LIVE dataset, we use linear  $\epsilon$ -SVR for regression with  $\lambda_1$  fixed to one. We use 1/4 of the training data as validation set. The minimal number of iterations of the optimization process is set to 50 and the maximal number of iterations is set to 250. After 50 iterations, if one of the two early stopping rules is true or the maximal number of iterations is reached, we terminate the filter training process. The learning rate in SGD is  $r_t = \frac{r_0}{\sqrt{1+t/N}}$ , where  $t$  is the number of iteration,  $N$  is the number of training samples and  $r_0$  is the initial learning rate.

For patch extraction, we fix the patch size  $BS = 5$  for all experiments. Given an image, non-overlapped patches are extracted and if the number of patches is more than 10000, we randomly sample 10000 patches of them.

For cross-validation, 80% of the data is used for training and the remaining 20% is used for testing. By default, given a fixed set of filters, CORNIA is trained using  $\nu$ -SVM with linear kernel<sup>2</sup>.

## Evaluation Metrics

For evaluating IQA measures, we follow common practice and use linear correlation coefficient (LCC) and Spearman rank order correlation (SROCC) as our evaluation measures. LCC is a measure of the prediction accuracy of a model and is defined as follows:

---

<sup>2</sup>LIBSVM parameters: '-s 4 -t 0 -c 1 -n 0.5'

$$LCC(X, Y) = \frac{\sum_{i=1}^n (X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^n (X_i - \mu_X)^2} \sqrt{\sum_{i=1}^n (Y_i - \mu_Y)^2}} \quad (2.12)$$

where  $\mu_X$  ( $\mu_Y$ ) is the average value of  $X_i$  ( $Y_i$ ).

SROCC is used to evaluate how well the relationship between the predicted quality score and true quality score can be described using a monotonic function. To compute the SROCC between two samples, we first convert the  $n$  raw scores  $X_i$  and  $Y_i$  to their ranks  $x_i$  and  $y_i$ , then  $SROCC(x, y)$  is given by

$$SROCC(x, y) = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^n (y_i - \mu_y)^2}} \quad (2.13)$$

where  $\mu_x$  ( $\mu_y$ ) is the average value of  $x_i$  ( $y_i$ ).

## 2.3.2 Experiments on Natural Scene Image

### 2.3.2.1 Dataset

Four IQA datasets were used to test CORNIA on natural scene images.

(1) *LIVE* [36]: the LIVE IQA dataset [36, 37] is perhaps the most widely used dataset for evaluating the performance of IQA algorithms. Images with five different types of distortions - JPEG2K compression noise(JP2K), JPEG compression noise (JPEG), white Gaussian noise (WN), Gaussian blurring (BLUR) and fast fading channel distortion (FF) which were derived from 29 non-distorted images are included. The *Differential Mean Opinion Score* (DMOS) associated with each distorted image is also provided. DMOS is generally in the range  $[0, 100]$ , where

lower DMOS indicates higher quality.

(2) *TID2008* [38]: The TID2008 dataset contains 25 reference images and 1700 distorted images from 17 different distortions at 4 levels each. In our work, we consider four of the 17 distortions which are most likely to occur in image processing systems – WN, JPEG, JP2K and BLUR. Each distorted image is associated with a Mean Opinion Score (MOS). Higher values of MOS (0 - minimal, 9 - maximal) correspond to higher visual quality of the image.

(3) CSIQ [39]: The CSIQ dataset consists of 30 reference images and their distorted versions with 6 different types of distortions at 4 to 5 different levels. For the CSIQ dataset, we consider the same four types of distortions – JP2K, JPEG, WN and BLUR. Each distorted image is associated with a DMOS score in the range  $[0, 1]$ .

(4) *LIVEMD* [40]: The LIVEMD (LIVE Multiply Distortion) dataset consists of images with multiple distortions. DMOS in the range  $[0, 100]$  is associated with each image. Two scenarios are considered:

- BLURJPEG: images are first blurred and then compressed by a JPEG encoder.
- BLURNOISE: images are first blurred and then corrupted by white Gaussian noise. For each type of distortion, 225 distorted images derived from 15 reference images are included.

In our experiments, human opinion scores from different datasets are all mapped to the range of  $[0, 100]$  as in the LIVE dataset.

### 2.3.2.2 Impact of unsupervised CORNIA parameters

A number of adjustable parameters have to be specified for the unsupervised CORNIA: (1) the number of patches extracted from each image; (2) the number of filters; (3) the size of raw image patch and (4) the encoding method. In this section, we focus on the effect of choosing different codebook sizes and encoders for the unsupervised CORNIA. Results reported in this section are obtained on the LIVE dataset.

**Effect of codebook size:** We consider codebook size of 50, 100, 200, 2500, 5000 and 10000. The median SROCC and LCC values from 100-fold cross-validation are presented in Fig. 2.5. We can see that the performance of unsupervised CORNIA improves as we increase the number of codewords in codebook and it drops significantly when the codebook size is smaller than 200. We will show later that the supervised CORNIA can boost the performance of small codebooks significantly.

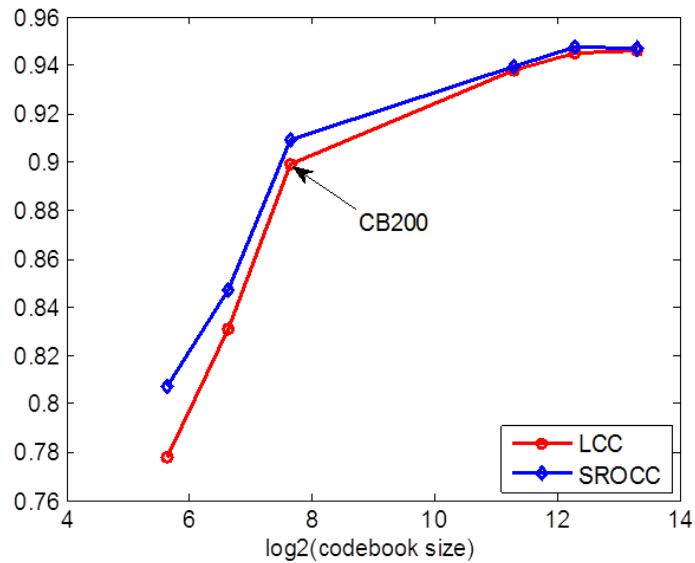


Figure 2.5: Effect of codebook size (tested on the LIVE database).



**Effect of encoding methods:** In addition to the soft-assignment (SA) encoding method, we also tried using sparse coding (SC) [31], locality-constrained linear coding (LLC) [32], “Localized” soft-assignment coding (LSA) [30] and conventional hard-assignment coding (HA) for encoding. SC and LSA were used with rectification and max pooling; LLC was used with max pooling but without rectification<sup>3</sup> and HA was used with average pooling and no rectification<sup>4</sup>. For LLC and LSA, we need to specify the number of nearest neighbors used for encoding, we used five for both. For all encoding methods, codebook of size 10000 is used and results are shown in Fig. 2.6. We found that the simple SA encoding slightly outperforms the other four encoding methods, even though LLC, LSA and SC have been shown to perform better than conventional SA in image classification problems.

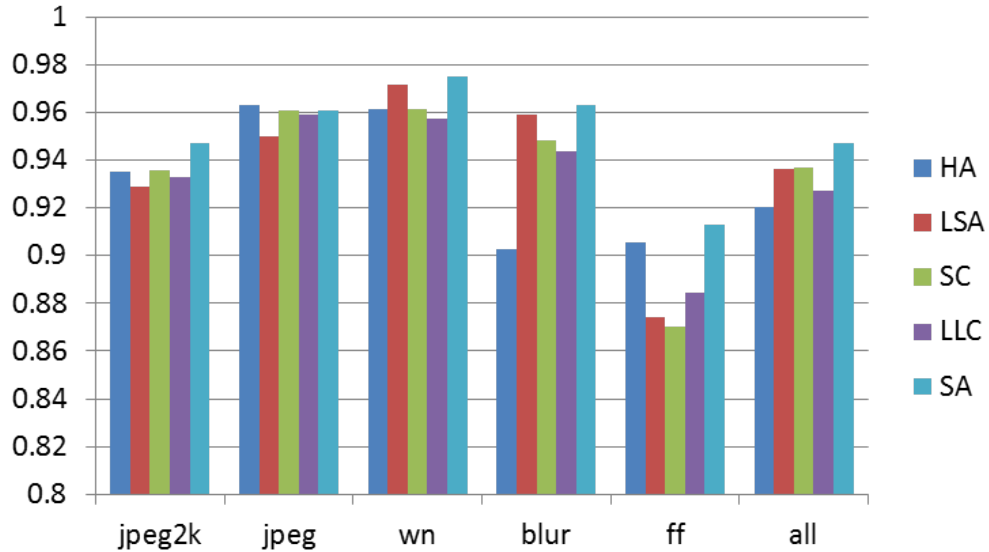


Figure 2.6: Effect of different encoders, evaluated using SROCC on the LIVE database.

<sup>3</sup>We found that using LLC without rectification performs better than with rectification.

<sup>4</sup>The way we compute HA code gives a vector with elements non-negative, thus no rectification can be performed

### 2.3.2.3 Evaluating unsupervised CORNIA

In the first set of experiments, we compare the unsupervised CORNIA with three state-of-the-art NR-IQA measures – DIIVINE [6], BLIINDS-II [7] and BRISQUE [8] and two FR measures – PSNR and SSIM.

#### **Results on the LIVE, TID2008, CSIQ and LIVEMD datasets**

The median value of LCC and SROCC from 1000 repeated experiments for four different datasets are shown in Tables 2.1, 2.2, 2.3 and 2.4 respectively. It is worth noting that DIIVINE, BLIINDS-II and BRISQUE were trained using SVM with RBF kernel while CORNIA was trained using linear SVM. The same set of training parameters are fixed for all experiments.

From these results we can see that

- Compared to state-of-the-art NR-IQA approaches, CB10000 is comparable to BRISQUE and on the LIVE dataset it outperforms DIIVINE and BLIINDS-II.
- CB10000 consistently outperforms PSNR on different datasets.
- CB10000 outperforms SSIM on the LIVE dataset and is comparable to SSIM on the other three datasets.

#### **Statistical Significance Test**

We performed a two sample T-test with 95% confidence level between SROCC generated by PSNR, SSIM, BRISQUE and our algorithms in 1000 iterations of experiments on the LIVE database. Test results are shown in Tables 2.5, 2.6, 2.7.

In these tables, **1** (**-1**) indicates the algorithm in the row is statistically superior

SROCC	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.872	0.885	0.941	0.764	0.875	0.867
SSIM	0.939	0.946	0.965	0.909	<b>0.941</b>	0.914
<i>DIIVINE</i>	0.913	0.910	<b>0.984</b>	0.921	0.863	0.916
<i>BLIINDS-II</i>	0.929	0.942	0.969	0.923	0.889	0.931
<i>BRISQUE</i>	0.914	<b>0.965</b>	0.979	0.951	0.877	0.940
<i>CB10000</i>	<b>0.947</b>	0.961	0.975	<b>0.963</b>	0.913	<b>0.947</b>
LCC	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.873	0.874	0.928	0.774	0.869	0.855
SSIM	0.920	0.955	0.982	0.891	<b>0.939</b>	0.906
<i>DIIVINE</i>	0.922	0.921	<b>0.988</b>	0.923	0.888	0.917
<i>BLIINDS-II</i>	0.935	0.968	0.980	0.938	0.896	0.930
<i>BRISQUE</i>	0.923	<b>0.973</b>	0.985	0.951	0.903	0.942
<i>CB10000</i>	<b>0.957</b>	<b>0.973</b>	0.986	<b>0.963</b>	0.925	<b>0.946</b>

Table 2.1: Evaluation on the LIVE distorted images. (*Italicized* algorithms are NR-IQA algorithms, others are FR-IQA algorithms.)

SROCC	JP2K	JPEG	WN	BLUR	ALL
PSNR	0.838	0.887	<b>0.917</b>	0.929	0.869
SSIM	<b>0.962</b>	<b>0.932</b>	0.847	<b>0.959</b>	<b>0.905</b>
<i>BRISQUE</i>	0.880	0.896	0.851	0.862	0.889
<i>CB10000</i>	0.937	0.926	0.911	0.931	0.899
LCC	JP2K	JPEG	WN	BLUR	ALL
PSNR	0.888	0.880	<b>0.945</b>	0.914	0.845
SSIM	<b>0.971</b>	<b>0.964</b>	0.816	<b>0.954</b>	0.902
<i>BRISQUE</i>	0.891	0.929	0.857	0.859	0.904
<i>CB10000</i>	0.945	0.963	0.903	0.927	<b>0.928</b>

Table 2.2: Evaluation on the TID2008 dataset. (*Italicized* algorithms are NR-IQA algorithms, others are FR-IQA algorithms.)

SROCC	JP2K	JPEG	WN	BLUR	ALL
PSNR	0.910	0.891	0.933	0.809	0.885
SSIM	<b>0.962</b>	<b>0.954</b>	0.912	<b>0.960</b>	<b>0.934</b>
<i>BRISQUE</i>	0.880	0.950	<b>0.956</b>	0.913	0.913
<i>CB10000</i>	0.925	0.912	0.952	0.935	0.905
LCC	JP2K	JPEG	WN	BLUR	ALL
PSNR	0.861	0.887	0.946	0.771	0.856
SSIM	0.906	<b>0.982</b>	0.910	0.945	<b>0.930</b>
<i>BRISQUE</i>	0.895	0.977	<b>0.964</b>	0.926	<b>0.930</b>
<i>CB10000</i>	<b>0.946</b>	0.962	0.952	<b>0.953</b>	0.928

Table 2.3: Evaluation on the CSIQ dataset. (*Italicized* algorithms are NR-IQA algorithms, others are FR-IQA algorithms.)

SROCC	BLURJPEG	BLURNOISE	ALL
PSNR	0.663	0.708	0.695
SSIM	0.905	<b>0.926</b>	<b>0.911</b>
<i>BRISQUE</i>	0.921	0.893	<b>0.911</b>
<i>CB10000</i>	<b>0.930</b>	0.915	<b>0.911</b>
LCC	BLURJPEG	BLURNOISE	ALL
PSNR	0.746	0.786	0.764
SSIM	0.943	<b>0.946</b>	<b>0.938</b>
<i>BRISQUE</i>	0.946	0.923	0.935
<i>CB10000</i>	<b>0.954</b>	0.942	0.936

Table 2.4: Evaluation on the LIVEMD dataset. (*Italicized* algorithms are NR-IQA algorithms, others are FR-IQA algorithms.)

	PSNR	SSIM	BRISQUE	CORNIA
PSNR	0	-1	-1	-1
SSIM	1	0	-1	-1
BRISQUE	1	1	0	-1
CORNIA	1	1	1	0

Table 2.5: Results of the two sample T-test performed between SROCC values obtained by different measures on the LIVE dataset.

	PSNR	SSIM	BRISQUE	CORNIA
PSNR	0	-1	-1	-1
SSIM	1	0	1	1
BRISQUE	1	-1	0	-1
CORNIA	1	-1	1	0

Table 2.6: Results of the two sample T-test performed between SROCC values obtained by different measures on the TID2008 dataset.

(inferior) than the algorithm in the column. 0 indicates the algorithm in the row is statistically equivalent to the algorithm in the column. We can see that our method offers the best performance on the LIVE dataset. We did not compare with DIIVINE and BLIINDS-II in this experiment, but in [8], BRISQUE is shown to be statistically better than DIIVINE and BLIINDS-II. Therefore, our method also outperforms DIIVINE and BLIINDS-II on the LIVE dataset. On the TID2008 dataset and the CSIQ dataset, the FR measure SSIM performs the best in terms of SROCC. On the TID2008 dataset, CORNIA slightly outperforms BRISQUE, while on the CSIQ dataset, BRISQUE slightly outperforms CORNIA.

### Dataset Independence Test

To show that our method does not depend on any particular dataset, we trained

	PSNR	SSIM	BRISQUE	CORNIA
PSNR	0	-1	-1	-1
SSIM	1	0	1	1
BRISQUE	1	-1	0	1
CORNIA	1	-1	-1	0

Table 2.7: Results of the two sample T-test performed between SROCC values obtained by different measures on the CSIQ dataset.

	CB10000	BRISQUE		CB10000	BRISQUE
SROCC	0.881	0.882	LCC	0.883	0.892

Table 2.8: Train on LIVE and test on TID2008

the prediction model on the LIVE dataset, and applied the trained model on the CSIQ and TID2008 datasets. The results are shown in Tables 2.8 and 2.9. We can see that CB10000 is comparable to BRISQUE in the dataset independence test.

#### 2.3.2.4 Evaluating the supervised CORNIA

In the second set of experiments, we evaluate the performance of the supervisedly learned filters.

##### **Results on the LIVE dataset**

For evaluation on the LIVE dataset, we split the dataset into 80% training set and 20% testing set. And in each iteration of the experiments, we train the codebook

	CB10000	BRISQUE		CB10000	BRISQUE
SROCC	0.899	0.899	LCC	0.914	0.927

Table 2.9: Train on LIVE and test on CSIQ

SROCC	JP2K	JPEG	WN	BLUR	FF	ALL
<i>CB10000</i>	<b>0.947</b>	<b>0.961</b>	<b>0.975</b>	<b>0.963</b>	<b>0.913</b>	<b>0.947</b>
<i>CB100</i>	0.915	0.846	0.953	0.946	0.878	0.839
<i>SF100</i>	0.924	0.928	0.962	0.961	0.879	0.920
LCC	JP2K	JPEG	WN	BLUR	FF	ALL
<i>CB10000</i>	<b>0.957</b>	<b>0.973</b>	<b>0.986</b>	<b>0.963</b>	<b>0.925</b>	<b>0.946</b>
<i>CB100</i>	0.918	0.843	0.970	0.947	0.878	0.821
<i>SF100</i>	0.929	0.940	0.978	0.960	0.888	0.921

Table 2.10: Comparing supervised and unsupervised CORNIA on the LIVE dataset.

and the prediction model on the training set, then test on the testing set. Since supervised fitler training process is time-consuming, this process is only repeated 100 times. Experimental results are presented in Table 2.10.

The parameters used in training superivsed CORNIA include:

- (1) For  $\epsilon$ -SVR:  $\epsilon = 1$ , cost  $C = 1$ , in Huber loss  $h = 0.9$ .
- (2) The balancing factor Eq. 2.4:  $\lambda_1 = 1$ .
- (3) For SGD: the initial learning rate  $r_0 = 0.001$ .

The optimization process on the entire LIVE from the first 51 iterations averaged over a 100-fold cross-validation experiment is shown in Fig. 2.7. It can be seen that the training process converges.

### Results on the CSIQ, TID2008 and LIVEMD datasets

For evaluation on the CSIQ, TID2008 and LIVEMD datasets, we train the supervised codebooks on distorted images with JP2K, JPEG, WN and BLUR distortions in the LIVE dataset. This process is repeated 10 times to obtain 10 codebooks. We select the codebook which performs the best on the validation set to extract features

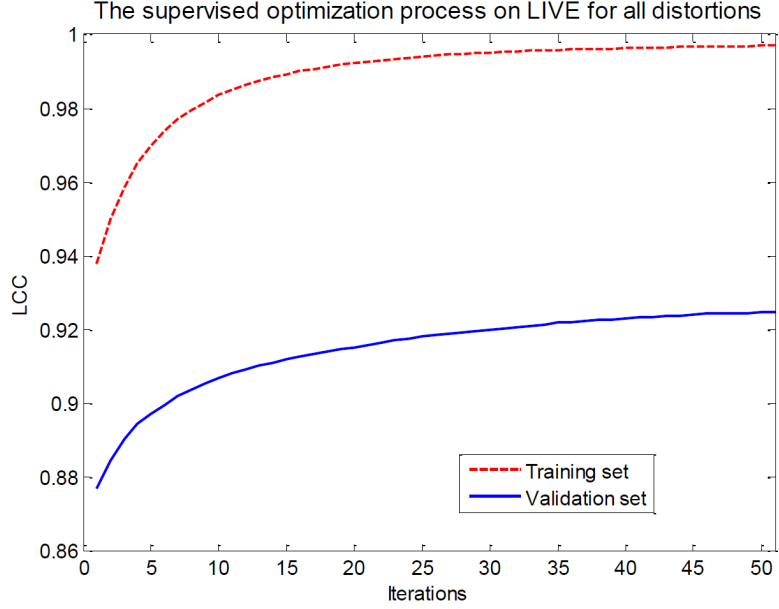


Figure 2.7: Optimization process of the first 51 iterations on training set and validation set (average LCC from 100 fold experiments on LIVE).

from images in the CSIQ, TID2008 and LIVEMD dataset. We then train a  $\mu$ -SVR with parameters specified in Section 2.3.1.

Experimental results are shown in Tables 2.11, 2.12 and 2.13. We can see that on the TID2008 and CSIQ dataset, significant performance gain is achieved by using supervised filter learning. SF200 is comparable to CB10000. However, on the LIVEMD dataset, the performance of SF200 is slightly worse than CB200. This is due to the fact that images in the LIVEMD dataset consist of multiple distortions, and we did not use examples with the same distortion to train the supervised codebook.

To visualize how the supervised filter learning changes the initial filters, we show the unsupervised codebook (CB200) and supervised codebook (SF200) filters in Fig. 2.8. We can see that after supervised training the filter pattern tends to be



SROCC	JP2K	JPEG	WN	BLUR	ALL
<i>CB10000</i>	0.937	0.926	<b>0.911</b>	<b>0.931</b>	<b>0.899</b>
<i>CB200</i>	0.932	0.899	0.872	0.889	0.826
<i>SF200</i>	<b>0.938</b>	<b>0.934</b>	0.893	0.901	0.898
<i>CB100</i>	0.928	0.870	0.853	0.889	0.808
<i>SF100</i>	0.926	0.907	0.848	0.861	0.865
LCC	JP2K	JPEG	WN	BLUR	ALL
<i>CB10000</i>	<b>0.945</b>	<b>0.963</b>	<b>0.903</b>	<b>0.927</b>	<b>0.928</b>
<i>CB200</i>	0.942	0.928	0.864	0.876	0.869
<i>SF200</i>	0.944	0.954	0.884	0.900	0.918
<i>CB100</i>	0.934	0.890	0.852	0.878	0.852
<i>SF100</i>	0.938	0.930	0.847	0.857	0.890

Table 2.11: Comparing supervised and unsupervised CORNIA on the TID2008 dataset.

SROCC	JP2K	JPEG	WN	BLUR	ALL
<i>CB10000</i>	0.925	0.912	<b>0.952</b>	<b>0.935</b>	<b>0.905</b>
<i>CB200</i>	0.913	0.880	0.933	0.921	0.867
<i>SF200</i>	<b>0.929</b>	<b>0.915</b>	0.921	0.918	0.904
<i>CB100</i>	0.911	0.865	0.931	0.920	0.849
<i>SF100</i>	0.916	0.901	0.932	0.919	0.907
LCC	JP2K	JPEG	WN	BLUR	ALL
<i>CB10000</i>	0.946	<b>0.962</b>	<b>0.952</b>	<b>0.953</b>	0.928
<i>CB200</i>	0.933	0.930	0.928	0.944	0.902
<i>SF200</i>	<b>0.950</b>	0.954	0.912	0.942	0.927
<i>CB100</i>	0.928	0.921	0.922	0.945	0.883
<i>SF100</i>	0.939	0.941	0.920	0.945	<b>0.930</b>

Table 2.12: Comparing supervised and unsupervised CORNIA on the CSIQ dataset.

SROCC	BLURJPEG	BLURNOISE	ALL
<i>CB10000</i>	<b>0.930</b>	<b>0.915</b>	<b>0.911</b>
<i>CB200</i>	0.903	0.900	0.892
<i>SF200</i>	0.920	0.897	0.884
<i>CB100</i>	0.895	0.909	0.876
<i>SF100</i>	0.906	0.893	0.886
LCC	BLURJPEG	BLURNOISE	ALL
<i>CB10000</i>	<b>0.954</b>	<b>0.942</b>	<b>0.936</b>
<i>CB200</i>	0.933	0.928	0.917
<i>SF200</i>	0.941	0.927	0.916
<i>CB100</i>	0.931	0.928	0.904
<i>SF100</i>	0.936	0.921	0.918

Table 2.13: Comparing supervised and unsupervised CORNIA on the LIVEMD dataset.

more random.

In the previous experiments, we used linear SVR with CORNIA. With the use of a small codebook, it is now possible to use a nonlinear SVR to boost the performance of CORNIA. Next, we test both unsupervised and supervised codebook using nonlinear  $\epsilon$ -SVR with RBF kernel<sup>5</sup>. For the TID2008 and CSIQ datasets, the use of RBF kernel can significantly boost the performance of both unsupervised and supervised CORNIA. *SF200-RBF* outperforms *CB10000* on TID2008 and CSIQ dataset. For the LIVEMD dataset, the RBF kernel does not seem to be helpful. Among all versions of CORNIA, we have tested, *CB10000* and *SF200-RBF* provide the best performance. Table 2.17 presents a comparison of the best performing CORNIA systems with previous methods (BRISQUE, PSNR and SSIM).

---

<sup>5</sup>Parameters in libsvm “-s 3 -p 1 -c 1024 -g 0.01”

SROCC	JP2K	JPEG	WN	BLUR	ALL
<i>CB10000</i>	<b>0.937</b>	0.926	<b>0.911</b>	<b>0.931</b>	0.899
<i>CB200-RBF</i>	0.931	0.910	0.851	0.911	0.894
<i>SF200-RBF</i>	0.932	<b>0.935</b>	0.886	0.911	<b>0.939</b>
<i>CB100-RBF</i>	0.913	0.884	0.868	0.887	0.834
<i>SF100-RBF</i>	0.928	0.923	0.876	0.889	0.880
LCC	JP2K	JPEG	WN	BLUR	ALL
<i>CB10000</i>	0.945	0.963	<b>0.903</b>	<b>0.927</b>	0.928
<i>CB200-RBF</i>	0.936	0.950	0.836	0.912	0.917
<i>SF200-RBF</i>	<b>0.946</b>	<b>0.965</b>	0.876	0.902	<b>0.947</b>
<i>CB100-RBF</i>	0.925	0.919	0.862	0.879	0.872
<i>SF100-RBF</i>	0.941	0.955	0.868	0.891	0.914

Table 2.14: Comparing supervised and unsupervised CORNIA with RBF kernel on the TID2008 dataset.

SROCC	JP2K	JPEG	WN	BLUR	ALL
<i>CB10000</i>	0.925	<b>0.912</b>	0.952	<b>0.935</b>	0.905
<i>CB200-RBF</i>	0.926	0.872	0.953	0.929	0.893
<i>SF200-RBF</i>	<b>0.933</b>	0.907	0.952	0.933	<b>0.923</b>
<i>CB100-RBF</i>	0.924	0.870	0.948	0.938	0.876
<i>SF100-RBF</i>	0.932	0.906	<b>0.958</b>	0.921	0.914
LCC	JP2K	JPEG	WN	BLUR	ALL
<i>CB10000</i>	0.946	<b>0.962</b>	0.952	0.953	0.928
<i>CB200-RBF</i>	0.946	0.930	<b>0.963</b>	0.951	0.918
<i>SF200-RBF</i>	<b>0.958</b>	0.953	0.962	0.953	<b>0.943</b>
<i>CB100-RBF</i>	0.941	0.920	0.951	<b>0.957</b>	0.898
<i>SF100-RBF</i>	0.953	0.944	0.961	0.945	0.931

Table 2.15: Comparing supervised and unsupervised CORNIA with RBF kernel on the CSIQ dataset.

SROCC	BLURJPEG	BLURNOISE	ALL
<i>CB10000</i>	<b>0.930</b>	<b>0.915</b>	<b>0.911</b>
<i>CB200-RBF</i>	0.894	0.887	0.888
<i>SF200-RBF</i>	0.908	0.893	0.895
<i>CB100-RBF</i>	0.843	0.857	0.850
<i>SF100-RBF</i>	0.866	0.849	0.848
LCC	BLURJPEG	BLURNOISE	ALL
<i>CB10000</i>	<b>0.954</b>	<b>0.942</b>	<b>0.936</b>
<i>CB200-RBF</i>	0.931	0.929	0.923
<i>SF200-RBF</i>	0.937	0.931	0.929
<i>CB100-RBF</i>	0.893	0.900	0.883
<i>SF100-RBF</i>	0.910	0.903	0.887

Table 2.16: Comparing supervised and unsupervised CORNIA with RBF kernel on the LIVEMD dataset.

SROCC	TID2008	CSIQ	LIVEMD
PSNR	0.869	0.885	0.695
SSIM	0.905	<b>0.934</b>	<b>0.911</b>
<i>BRISQUE</i>	0.889	0.913	<b>0.911</b>
<i>CB10000</i>	0.899	0.905	<b>0.911</b>
<i>SF200-RBF</i>	<b>0.939</b>	0.923	0.895
LCC	TID2008	CSIQ	LIVEMD
PSNR	0.845	0.856	0.764
SSIM	0.902	0.930	<b>0.938</b>
<i>BRISQUE</i>	0.904	0.930	0.935
<i>CB10000</i>	0.928	0.928	0.936
<i>SF200-RBF</i>	<b>0.947</b>	<b>0.943</b>	0.929

Table 2.17: Comparing CORNIA with PSNR, SSIM and BRISQUE on TID2008, CSIQ and LIVEMD datasets.



(a)



(b)

Figure 2.8: (a) 200  $5 \times 5$  filters in CB200. (b) 200  $5 \times 5$  filters SF200.

### 2.3.2.5 Evaluating Speed

Since IQA measures are often used in real-time imaging or communication systems, speed is an important factor determining whether an IQA measure can be used. We tested the speed of the supervised method with 100 and 200 filters, the unsupervised method with 10000 filters and three other recent NR-IQA measures.  $512 \times 768$  image is used in the test. All three methods are implemented in Matlab

	<i>CORNIA-100</i>	<i>CORNIA-200</i>	BRISQUE	<i>CORNIA-10000</i>	BLIINDS-II	DIIVINE
Time	0.027	0.033	0.112	0.352	81.08	29.88

Table 2.18: Feature extraction time (in seconds).

and are tested on a SunFire x4170 with 2.80GH processor. We consider only feature extraction time and results are shown in Table 2.18. It is clear that CORNIA with a small codebook of size 100 or 200 is much faster than other NR-IQA methods. Suppose the image size is  $w \times h$ , codebook size is  $K$ , patch size is  $BS$  and for each image we extract  $N$  patches, then the computational complexity of our method is  $O(K \times N \times BS^2)$ .

### 2.3.3 Experiments on Document Images

Next, we present experimental results on document images. Instead of predicting human perceived image quality, for document IQA (Doc-IQA), we are interested in predicting the OCR accuracy with respect to a specific OCR software, this has been shown to be useful in many document image applications.

#### **Dataset**

To test our method on document images, we use the *Sharpness-OCR-Correlation*(SOC) dataset [41]. The SOC dataset contains camera-captured documents with blur. Twenty-five (25) non-distorted images in this dataset are taken from two freely available datasets - *University of Washington Dataset* [42] and *Tobacco Database* [43]. For each document, multiple photos were taken from a fixed distance to capture the whole document, but the camera was focused at varying distance to generate

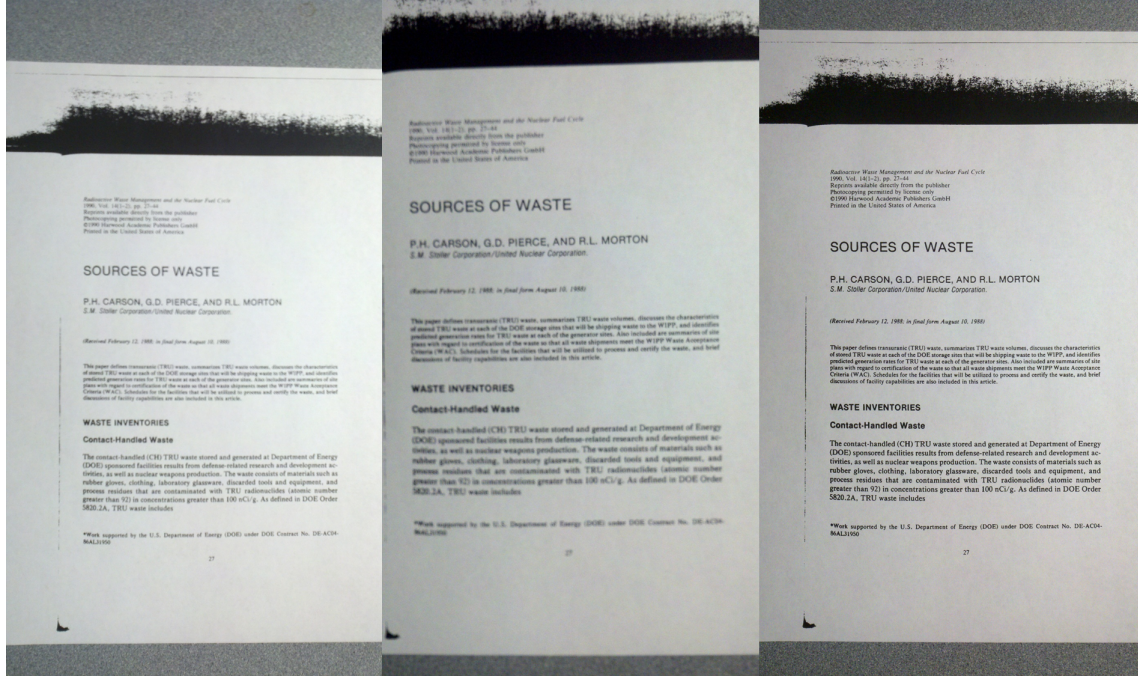


Figure 2.9: Examples of images in the SOC dataset.

a series of images with focal blur. A total of 25 such sets, each consisting of 6-8 high-resolution images (dimension:  $3264 \times 1840$ ) were created using an Android phone with an 8 mega-pixel camera. ABBYY Fine Reader [44] was used to obtain the OCR results and OCR accuracy were computed using ISRI-OCR evaluation tool [45]. OCR accuracy is in the range from 0 to 1. Examples of images from this dataset are shown in Fig.2.9. These images are not well aligned with each other. Positioning and lighting conditions for images captured for the same document may be different. As is shown in Fig.2.10, the distribution of OCR accuracy values for images in this dataset is highly imbalanced.

## Tests on the SOC dataset

Experimental results on the SOC dataset are shown in Table 2.19. Only one type of distortion is included in the dataset, therefore CB10000 consists of highly redundant

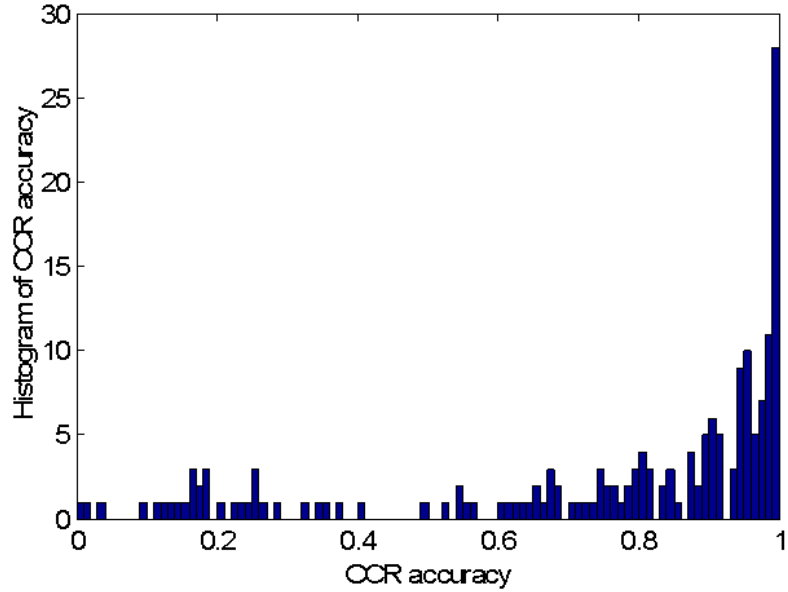


Figure 2.10: Distribution of OCR accuracy values of images in the SOC dataset.

features and only a very small set of discriminative features is critical to the success of the algorithm. It is important to include filters corresponding to these features in the codebook. We can see from the results that increasing codebook size does not necessarily improve the performance for unsupervised CORNIA. Since a fairly small codebook (e.g. *CB50*) performs comparable to a large codebook (*CB10000*), it is unnecessary to apply supervised filter learning to further reduce the codebook size. We therefore did not test supervised CORNIA for this task.

Consider two document images with the same level of blur distortion. They are expected to have similar OCR accuracy. However, it is hard to tell which image has higher OCR accuracy based on their visual appearances, since OCR accuracy is also related to the content of the document. This explains why we obtained LCC around 90%, while SROCC, which measures the monotonicity of the prediction



	BRISQUE	<i>CB10</i>	<i>CB50</i>	<i>CB10000</i>
SROCC	0.814	0.818	<b>0.850</b>	0.840
LCC	0.926	0.905	0.925	<b>0.927</b>

Table 2.19: Median LCC and SROCC with 1000 iterations of experiments on the SOC dataset.

model is much lower than the corresponding LCC.

## 2.4 Discussion and Open Problems

### 2.4.1 Discussion

Several issues should be more fully considered regarding the final performance of our approach. For supervised filter learning, the reported results were obtained using SGD with 1 epoch. We have also tested using SGD with 5 and 10 epochs, which converge faster than the 1-epoch SGD. However, the overall training time does not decrease since each iteration will take longer with 5-epoch training. Similarly the overall performance also does not vary much. To further improve the performance, we can repeat the iterative optimization process several times and choose the final learned model as the one which gives the best performance on the validation set.

When the initial set of filters obtained in an unsupervised setting captures the distortion properties well, supervised learning may not be very helpful. Using our approach, one can decide whether a supervised extension is needed based on specific prediction tasks. Boosting over a large set of filters, say 10000, will not lead to significant performance improvement and the over-fitting problem may be more severe. However, monitoring the training process on a validation set ensures that

our method will not lead to a decrease in the overall performance.

We can see from the experiments that when testing on one type of distortion, a relatively small codebook is enough to achieve good results. The advantage of using a large codebook is more obvious when testing on several different types of distortions. In practice, when only one type of distortion is of interest, we may try unsupervised CORNIA with a small codebook first. If the small codebook does not yield good results, we may then try larger ones.

### 2.4.2 Open Problems

There are a number of open problems and our CORNIA framework can be extended in the following ways:

#### **1. Convolutional neural networks**

Supervised CORNIA employs a two-layer structure which learns the filters and weights in the regression model simultaneously based on an EM like approach. This structure can be viewed as an empirical implementation of a two layer neural network. However, it has not utilized the full power of neural networks. One option would be to apply convolutional neural network to raw image pixels to learn discriminative features in a more unified way.

#### **2. Applications in Video Quality Assessment (VQA)**

Video Quality Assessment (VQA) shares many of the same challenges as IQA, so in theory CORNIA could be applied to such problems. One intuitive approach would be applying CORNIA to each video frame, then pooling through the whole video to

obtain its quality score. Since we usually only have the quality score for the video instead of each frame, one may train CORNIA on synthetic scores (See Chapter 3) for each frame. Then various pooling methods, such as the hysteresis temporal pooling [46], could be applied to obtain the overall video quality. In this case, no human opinion scores are involved in the VQA process (only synthetic scores are used). One may use the histogram of the frame score (or frame difference scores) or other informative statistics to form a feature vector for the entire video, then use the human opinion scores to train a prediction model.

A second approach to adapt CORNIA to the VQA problem may be to replace the 2D filters with 3D filters. Each 3D filter extracts features on a 3D cube in a video. Therefore temporal information can be incorporated in the 3D CORNIA feature.

### **3. Quality assessment for face recognition in videos**

In video based face recognition, face images are typically captured in unconstrained environments where illumination, facial expression, pose, resolution, scale, alignment, occlusions, shadowing, motion blur and the change of focus may vary across the video sequence.

The task of VQA for Face recognition (F-VQA) is very different from conventional IQA. First, conventional IQA methods focus primarily on quality variations due to external degradations. In these cases, distorted images under investigation often suffer only one primary type of distortion, which is uniformly distributed over the entire image. However, the inherent properties of video capture suggest that video frames usually suffer multiple types of distortions, and the distribution of the

distortion may not be uniform. Second, F-VQA should take into consideration face specific qualities such as face geometry, pose, eye detectability, illumination angles, while these factors are not considered in conventional IQA systems.

The success of CORNIA for IQA motivates the use of feature learning for F-VQA, where designing hand-craft features is even more difficult and less effective than conventional IQA due to the complexity of the video capturing process and the face specific properties. Previous approaches for F-VQA perform feature extraction and feature fusion independently. By applying the supervised feature learning, we can have a unified process for simultaneously learning discriminative features and prediction (regression or classification) models.

#### **4. Local Quality Estimation**

In many IQA applications, the distribution of noise may not be uniform. For example, packet loss in wireless communication channels may lead to local distortions that only affect relatively small image regions. It is therefore important to develop a method that can be applied to local image regions. To approach this problem, an intuitive extension for CORNIA is to divide one image into several regions, then apply CORNIA on each image region independently. Finally, the average score of each region can be considered as the whole image quality score.

## Chapter 3: Blind Learning of Image Quality based on Synthetic Scores

### 3.1 Introduction

State-of-the-art general purpose NR-IQA methods [6–8,19,20] rely on examples of distorted images and corresponding human opinion scores to learn a regression function that maps image features to quality scores. This type of model is considered “opinion-aware” (OA) because human opinion scores are provided for the distorted images. A large set of training images with scores is required to train a reliable OA-NRIQA model, but obtaining opinion scores can be time-consuming and expensive. To overcome this limitation, there has been an increasing interest in learning “opinion-free” (OF) NR-IQA models [47–49], which do not require human opinion scores for training.

We developed a simple yet effective method for extending OA-NRIQA models to OF-NRIQA models, which we refer to as BLISS (Blind Learning of Image quality based on Synthetic Scores). Instead of training on human opinion scores, we train NR-IQA models on full-reference (FR) image quality measures. FR measures directly quantify the differences between distorted images and their undistorted reference images and are easy to obtain. State-of-the-art FR measures yield high correlation with human opinion scores, so they can be used for training NR-IQA

models. Different FR measures may quantify visual quality in different ways and no single method typically gives the best performance in all situations. We apply unsupervised rank aggregation to combine different FR measures for generating a better baseline with which to train. Extensive experiments on three standard IQA datasets show that the our method significantly outperforms previous OF-NRIQA methods. Furthermore, models trained on the synthetic scores (including FR measures and combined synthetic scores) are comparable to models trained on the human opinion scores. This observation implies that we may replace human opinion scores by the synthetic scores in training NR-IQA models without performance loss. The strategy of training on synthetic scores helps to overcome the bottleneck arising from limited training data due to lack of expensive human opinion scores and allows to use a large set of data for training.

In this work, our contributions are two-fold. First, we show that FR measures can be used to replace human opinion scores for training NR-IQA models. This is an extremely flexible strategy and can be used with any well established NR-IQA model. Second, we develop an effective method to combine FR measures in an unsupervised way. The combined synthetic scores yield high correlation with human opinion scores and outperform each individual FR measure anticipated in the combination.

In the remainder of this chapter, we first briefly review previous work on OF-NRIQA model and FR measure combination in Section 3.2. We then describe our unsupervised rank aggregation based method for FR measure combination in Section 3.3. Experimental results are presented in Section 3.4. Section 3.5 concludes our

work and presents extensions to our current work.

## 3.2 Related Work

### 3.2.1 OF-NRIQA Models

The first OF-NRIQA model published in the literature was the TMIQ model introduced by Mittal et al. [47]. TMIQ applies probabilistic latent semantic analysis (pLSA) to quality-aware visual words extracted from a large set of pristine and distorted images to uncover latent characteristics or “topics” that are essential for visual quality. The topic mixing coefficients are estimated for the pristine images. Then, given a test image, its estimated topic mixing coefficients are compared to those for the pristine images and their differences are used to infer the quality for the test image. This method suffers from poor performance compared to state-of-the-art OA-NRIQA models.

Later Mittal et al. introduced another OF-NRIQA model – NIQE [48]. NIQE builds a multivariate Gaussian (MVG) model for the natural scene statistic (NSS) features of sharp image regions extracted from pristine images. For a test image, the distance between the MVG constructed from the NSS features of the test image and the MVG model constructed from pristine images is computed as the quality measure. NIQE significantly outperforms TMIQ, yet it does not require distorted images for training and thus is “distortion unaware” and “completely blind”. NIQE was shown to perform well on the five types of distortions in the LIVE dataset. This method however may not work universally well on all types of distortions, and when

it fails, it is hard to adjust the model to improve the performance since the model does not incorporate examples of distorted images in the training.

Xue et al. proposed a quality-aware clustering (QAC) method [49] for OF-NRIQA. QAC assigns each image patch a quality score based on a FR measure, then applies clustering to patches at different quality levels. Each cluster centroid is associated with a quality score. For a test image, overlapped patches are extracted, and each patch is compared to the quality aware cluster centroids. The quality score of its nearest neighbor is assigned to the patch. The final quality score for the test image is the weighted average of the patch level quality score.

None of these previous OF-NRIQA models apply discriminative training in constructing the model. However, the best performance of previous OA-NRIQA models are usually achieved by discriminative training (for example, Support Vector Regression (SVR)). The current OF-NRIQA models are all inferior to state-of-the-art OA-NRIQA models. Our method can be applied to extend existing OA-NRIQA models to OF-NRIQA models and achieve comparable performance.

### 3.2.2 Combining Multiple Full-Reference Measures

FR-measure combination methods aim to combine multiple types of FR-measures to yield a better quality measure [50]. Different FR measures quantify visual quality from different aspects and typically no single method gives the best performance in all situations. Therefore combining multiple FR measures may produce a better IQA measure which outperforms individual FR measures anticipated in the combination.



Liu et al. [50] introduced a supervised FR measure combination method. A nonlinear regression function is learned to map a feature vector formed by multiple FR measures to a human opinion score. This method requires human opinion scores to train the regression model and thus is not suitable to use in the “opinion-free” scenario.

We develop an unsupervised FR measure combination method. Given a set of images, we first apply unsupervised rank aggregation to obtain a single consensus ranking based on multiple FR measures. We then adjust a given FR measure based on the consensus ranking to generate combined synthetic scores. To the best of our knowledge, this is the first work that approaches the FR measure combination problem in an unsupervised way.

### 3.3 Unsupervised FR measure combination

In this section, we describe our unsupervised FR measure combination method, which involves two steps: generating consensus ranking and score adjustment. The method is summarized in Fig. 3.1.

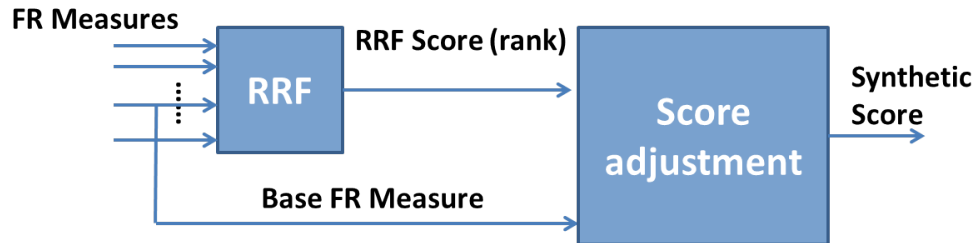


Figure 3.1: Overview of the unsupervised FR measure combination method.

### 3.3.1 Full-Reference Measures

FR-IQA measures are computed based on the differences between distorted images and their undistorted reference images. Five different FR measures are used in our experiments: GMSD [51], VIF [52], FSIM, FSIMC [53] and WSSIM [54].

- **GMSD** (Gradient Magnitude Similarity Deviation) computes the pixel-wise gradient magnitude similarity (GMS) and uses this to generate the final measure as the standard deviation of the GMS map.
- **VIF** (Visual Information Fidelity Index) models image sources using a wavelet domain Gaussian Scale Mixture (GSM) model. It measures the information shared between the source image and the distorted image based on an image distortion channel and a visual distortion model.
- **FSIM** (Feature Similarity Index) is based on low-level features, where phase congruency is used as the primary feature. **FSIMC** is FSIM applied to color images.
- **IW-SSIM** (Information content Weighted SSIM) is an enhanced version of the Structural Similarity Index (SSIM) [55] measure, that performs pooling over the SSIM map using weights that are proportional to the local information content.

These five FR measures achieve state-of-the-art performance on standard IQA datasets, so we select them to use in our system.

### 3.3.2 Combining Full-Reference Measures

Suppose we have  $K$  different FR measures and  $N$  training images  $I_i, i = 1, \dots, N$ . The first step in combining these FR measures is to construct a consensus ranking via unsupervised rank aggregation. The rank aggregation problem is crucial for many applications such as meta-search, crowdsourcing and social choice. There are many well established methods for this problem. We have selected and used the *Reciprocal Rank Fusion* (RRF) to generate the consensus ranking.

RRF [56] was initially proposed for combining the document rankings from multiple Information Retrieval (IR) systems. Despite its simplicity, RRF is one of the top performing unsupervised rank aggregation methods. According to a recent study in [57], RRF outperforms other competing unsupervised rank aggregation methods on the LETOR4.0 [58] datasets. The RRF score for image  $I_i$  is given by

$$RRFscore(I_i) = \sum_{k=1}^K \frac{1}{\gamma + r_k(i)} \quad (3.1)$$

where  $r_k(i)$  is the rank of  $I_i$  given by the  $k$ -th FR measure and  $\gamma = 60$  is a constant. According to [56], the constant  $\gamma$  mitigates the impact of high rankings by outlier systems. The rank of  $I_i$  given by the  $RRFscore(I_i)$  is denoted as  $t_i$ .

The pair-wise spearman rank order correlation coefficient (SROCC) between FR measures and RRF scores on the LIVE dataset are shown in Table 3.1. For VIF, FSIM, FSIMC and WSSIM, a higher value indicates good quality while for GMSD a smaller value indicates good quality. To compute the SROCC in Table 3.1, we use

SROCC	GMSD	VIF	FSIM	FSIMC	WSSIM	RRF
GMSD	–	0.9637	0.9896	0.9908	0.9769	0.9911
VIF	0.9637	–	0.9765	0.9745	0.9781	0.9844
FSIM	0.9896	0.9765	–	0.9986	0.9902	0.9978
FSIMC	0.9908	0.9745	0.9986	–	0.9897	0.9976
WSSIM	0.9769	0.9781	0.9902	0.9897	–	0.9939
RRF	0.9911	0.9844	0.9978	0.9976	0.9939	–

Table 3.1: Pair-wise SROCC between FR measures and RRFscore (Evaluated on LIVE).

the negative RMSD value. It can be seen that the RRF score has a high SROCC for each individual FR measure. This means that the rank given by the RRF is consistent with all five FR measures.

We note that the RRF score cannot be directly used as image quality score, because the  $RRFscore(I_i)$  is a quality indicator of  $I_i$  relative to other images in the dataset. It does not directly reflect the quality of  $I_i$ . In order to generate a valid quality measure, we adjust the score of a base FR measure according to the RRF rank. Suppose the score of a base FR measure for  $I_i$  is  $y_i$  and a higher score indicates better quality and a smaller rank  $t_i$ . The final combined quality score  $s$  is obtained by minimizing the following objective function.

$$\begin{aligned}
L(s) = & \sum_{i=1}^N (s_i - y_i)^2 + \\
& \lambda \sum_{i < j} (s_i - s_j) \mathbf{1}(t_i > t_j) + (s_j - s_i) \mathbf{1}(t_i < t_j)
\end{aligned} \tag{3.2}$$

where  $\lambda = \frac{(max(y)-min(y))\lambda_0}{N}$  is a constant balancing factor and  $\mathbf{1}(x) = 1$  if  $x$  is true,  $\mathbf{1}(x) = 0$  otherwise. The first term in the above equation tends to minimize the mean squared error between  $s$  and  $y$ . The second term penalizes the inconsistency

of pair-wise preferences between  $s$  and  $t$ . An optimal  $s$  can be found by setting the derivative of  $L(s)$  with respect to  $s$  equal to 0, which yields a simple closed form solution as follows:

$$\frac{\partial L(s)}{\partial s} = 0 \Rightarrow s_i = y_i - \frac{(\max(y) - \min(y))\lambda_0}{2N}n_i \quad (3.3)$$

where  $n_i = |\{j : t_j < t_i\}| - |\{j : t_j > t_i\}|$ .

The success of the combination method relies on the uniformity of the score distribution. This condition is a typical property of image quality datasets, since if the quality distribution is imbalanced, it would not be a good benchmark for evaluating IQA systems and it would be hard to use the dataset to train any NR-IQA models to achieve good performance. Our method works the best when all FR measures involved in the combination have similar performance. We can compute the pair-wise SROCC between different FR measures, and FR measures that have low correlation with other measures may be removed. In Section 3.4, we show experimentally that the synthetic scores defined in Eq. 3.3 have higher correlation with human opinion scores than their base FR measures on multiple IQA datasets.

### 3.3.3 Computational Complexity

Sorting  $N$  real numbers using merge sort or quick sort has time complexity  $O(N\log(N))$ . To compute the *RRF* scores given  $K$  different FR measures, we have to sort the  $N$  images  $K$  times according to different measures, therefore the complexity is  $O(KN\log(N))$  (usually  $K \ll N$ ). And the complexity for computing

$n_i$  in Eq. 3.3 is  $O(N\log(N))$ . Therefore the overall complexity is  $O(KN\log(N))$ .

### 3.3.4 Training

Once the combined synthetic scores are computed, we can use them to replace human opinion scores for training a NR-IQA model, and the original “opinion-aware” models will become “opinion-free”. For training NR-IQA models, we use Support Vector Regression (SVR). The computational complexity for predicting the quality of a new image is determined by the base NR-IQA model. No additional overhead will be introduced by BLISS. We can also use a single FR measure for training, but as will be shown later, moderate performance improvements can be achieved by training on combined synthetic scores.

## 3.4 Experiments

In this section, we present experiments on three standard IQA benchmarks to demonstrate the effectiveness of our method.

### 3.4.1 Experimental Protocol

#### Datasets

Three IQA databases (also described in Chapter 2) were used in our experiments:

- (1) LIVE [36]: The LIVE dataset contains a total of 779 distorted images derived from 29 reference images. Each reference image is distorted by five different distortions – JP2k compression (JP2K), JPEG compression (JPEG), White Gaus-

sian (WN), Gaussian blur (BLUR) and Fast Fading (FF) at 7-8 different levels. Note that among the 982 images in the LIVE dataset, only 779 of them are distorted images. The same reference image may occur multiple times in the dataset. Therefore the correlation coefficients should be computed only on distorted images to truly reflect the performance of the algorithm.

(2) TID2008 [38]: The TID2008 dataset contains 1700 distorted images derived from 25 reference images. A total of 17 different distortions at four degradation levels are included in this dataset. In our experiments, we only examine the four common distortions that are shared by the LIVE dataset, i.e. JP2K, JPEG, WN and BLUR.

(3) CSIQ [39]: The CSIQ dataset consists of 30 reference images and their distorted versions with 6 different types of distortions at 4 to 5 different levels. For the CSIQ dataset, we consider the same four types of distortions – JP2K, JPEG, WN and BLUR.

## Evaluation

The performances of IQA measures are evaluated using Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC). It is a common practice to evaluate the FR-IQA measures with a curve-fitting procedure [59], since different IQA measures may not lie in the same range. A similar procedure may also be applied to OF-NRIQA models. A logistic regression function,  $Q_p = \beta_1(\frac{1}{2} - \frac{1}{\exp(\beta_2(Q - \beta_3))}) + \beta_4Q + \beta_5$ , is used to map the original IQA measures to the range of human opinion scores. In our experiments, we randomly select 80% of the reference images and their associated distorted versions for training to obtain  $\beta_i, i = 1, \dots, 5$  and use the remaining 20% of the reference images and their associ-

ated distorted versions for testing. This procedure is repeated 1000 times and the median values of LCC and SROCC are reported.

### 3.4.2 Implementation Details

**Training Set Construction:** We downloaded 100 high resolution images under the Attribution License from flickr.com. The topics of these images include animal, building, indoor scene, forest, human, plant, man-made object, food, sports, etc. From each image, we derive one non-distorted image and the 100 images form our reference image set. Then from each reference image, we generate distorted images with four types of distortions including JPEG and JPEG2k compression, white Gaussian noise and Gaussian blurring. For each distortion, 8 distortion levels are considered. A total of 3300 images are generated to form our training set of 3200 distorted images and the 100 reference images. FR measures and combined synthetic scores are computed as groundtruth for the training set. These scores are mapped to the range of  $[0, 100]$  by a linear function to make it consistent with the range of DMOS in the LIVE dataset. The quality scores of reference images are set to  $-1$ .

**Base NR-IQA model:** We use CORNIA [19] as the base NR-IQA method because it gives state-of-the-art performance with the use of a linear regression function. Our training set contains 3300 images and training a nonlinear SVR would be time consuming. To speed up the training process, we use the fast liblinear library [60].

**Base FR measure:** Among the five types of FR measures anticipated in the score combination, we select GMSD as the base FR measure because GMSD [51] yields



$BS$	$cbsize$	$C$	$\epsilon$	$\lambda_0$
5	10000	100	1	4

Table 3.2: Parameters used in our experiments.

high linear correlation with human opinion scores without applying any nonlinear fitting and it is very efficient to compute.

**Parameters:** Several parameters have to be specified for our experiments. (1) In CORNIA feature extraction:  $BS$  - patch size;  $cbsize$  - codebook size and (2) in learning the regression function using liblinear:  $C$  - cost in the loss function;  $\epsilon$  - parameter in  $\epsilon$ -insensitive loss function used in  $\epsilon$ -SVR. The solver we used in liblinear is the L2-regularized L2-loss support vector regression (primal). (3)  $\lambda_0$  the balancing factor in Eq. 3.2. Table 3.2 shows the values of these parameters.

### 3.4.3 Evaluation

Next, we test our method on the three standard IQA benchmarks with experimental protocols described above.

#### 3.4.3.1 Comparison with FR and OF-NRIQA Algorithms

We first compare our method with previous methods including the FR measures: PSNR and SSIM [55] and state-of-the-art OF-NRIQA methods: QAC [49] and NIQE [48]. We test on the four types of distortions which are shared by the LIVE, CISQ and TID2008 datasets (JPEG2K, JPEG, WN and Gaussian BLUR). BLISS is trained on the entire flickr dataset. Test results on each subset and all

four subsets combined are reported. BLISS-S is trained using GMSD, which yields the best performance among all the five FR measures. BLISS-C is trained on the combination of five FR measures using Eq.3.3. It is worth noting that the same parameters specified in Table 3.2 are used for experiments on all three datasets.

Results on the LIVE dataset, the CSIQ dataset and the TID2008 dataset are presented in Tables 3.3, 3.4 and 3.5 respectively. We can see that BLISS significantly outperforms the other two competing OF-NRIQA models. BLISS-C slightly outperforms BLISS-S. Tables 3.6, 3.7, 3.8 show results of two sample T-test with 5% significance level on SROCC values obtained from the LIVE dataset, where 1 (-1) implies the algorithm in the row is statistically superior (inferior) to the algorithm in the column. 0 indicates the algorithm in the row is statistically equivalent to the algorithm in the column. We can see that the combined score for training outperforms the use of a single FR measure on the LIVE and CSIQ datasets and they are comparable on the TID2008 dataset.

The standard deviations (STD) of the SROCC and LCC obtained from a 1000-fold cross-validation experiment on the LIVE dataset are shown in Table 3.9. We can see that BLISS-C and BLISS-S tend to have smaller STD compared to other methods. This demonstrates the consistency of our method.

### 3.4.3.2 Comparison with OA-NRIQA Algorithms

In the second set of experiments, we use human opinion scores (DMOS) and synthetic scores (SS) to train two state-of-the-art BIQA models BRISQE [8] and

SROCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.870	0.885	0.942	0.761	0.867
SSIM	0.939	0.946	0.964	0.907	0.910
NIQE	0.924	0.944	<b>0.972</b>	0.939	0.922
QAC	0.868	0.938	0.952	0.918	0.877
BLISS-S	0.911	0.935	0.965	0.954	0.935
BLISS-C	<b>0.928</b>	<b>0.946</b>	0.970	<b>0.959</b>	<b>0.943</b>
LCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.873	0.876	0.926	0.766	0.853
SSIM	0.921	0.955	0.982	0.891	0.900
NIQE	0.931	0.957	0.955	0.950	0.919
QAC	0.851	0.943	0.924	0.919	0.863
BLISS-S	0.911	0.958	0.974	0.958	0.933
BLISS-C	<b>0.933</b>	<b>0.965</b>	<b>0.976</b>	<b>0.967</b>	<b>0.939</b>

Table 3.3: Results on LIVE.

SROCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.910	0.891	0.933	0.809	0.885
SSIM	0.962	0.954	0.912	0.960	0.934
NIQE	0.925	0.883	0.835	0.907	0.887
QAC	0.888	<b>0.912</b>	<b>0.865</b>	0.852	0.858
BLISS-S	0.935	0.889	0.815	0.913	0.899
BLISS-C	<b>0.949</b>	0.910	0.848	<b>0.917</b>	<b>0.918</b>
LCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.861	0.887	0.946	0.771	0.856
SSIM	0.906	0.982	0.910	0.945	0.930
NIQE	0.934	0.945	0.834	0.929	0.904
QAC	0.896	0.947	<b>0.911</b>	0.861	0.890
BLISS-S	0.951	0.952	0.833	0.944	0.927
BLISS-C	<b>0.965</b>	<b>0.959</b>	0.863	<b>0.945</b>	<b>0.938</b>

Table 3.4: Results on CSIQ.

SROCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.838	0.887	0.917	0.929	0.869
SSIM	0.962	0.932	0.847	0.959	0.905
NIQE	0.887	0.875	<b>0.817</b>	0.845	0.795
QAC	0.890	0.887	0.717	0.856	0.861
BLISS-S	0.919	0.922	0.779	0.869	0.898
BLISS-C	<b>0.923</b>	<b>0.926</b>	0.807	<b>0.880</b>	<b>0.899</b>
LCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.888	0.880	0.945	0.914	0.845
SSIM	0.971	0.964	0.816	0.954	0.902
NIQE	0.911	0.921	<b>0.796</b>	0.849	0.804
QAC	0.878	0.917	0.736	0.842	0.842
BLISS-S	<b>0.945</b>	<b>0.955</b>	0.748	0.875	0.910
BLISS-C	0.941	0.952	0.770	<b>0.880</b>	<b>0.917</b>

Table 3.5: Results on TID2008.

SROCC	PSNR	SSIM	NIQE	QAC	BLISS-S	BLISS-C
PSNR	0	-1	-1	-1	-1	-1
SSIM	1	0	-1	1	-1	-1
NIQE	1	1	0	1	-1	-1
QAC	1	-1	-1	0	-1	-1
BLISS-S	1	1	1	1	0	-1
BLISS-C	1	1	1	1	1	0

Table 3.6: Results of the two sample T-test performed between SROCC values obtained by different measures (Evaluated on LIVE).

SROCC	PSNR	SSIM	NIQE	QAC	BLISS-S	BLISS-C
PSNR	0	-1	1	1	-1	-1
SSIM	1	0	1	1	1	1
NIQE	-1	-1	0	-1	-1	-1
QAC	-1	-1	1	0	-1	-1
BLISS-S	1	-1	1	1	0	0
BLISS-C	1	-1	1	1	0	0

Table 3.7: Results of the two sample T-test performed between SROCC values obtained by different measures (Evaluated on TID2008).

SROCC	PSNR	SSIM	NIQE	QAC	BLISS-S	BLISS-C
PSNR	0	-1	-1	1	-1	-1
SSIM	1	0	1	1	1	1
NIQE	1	-1	0	1	-1	-1
QAC	-1	-1	-1	0	-1	-1
BLISS-S	1	-1	1	1	0	-1
BLISS-C	1	-1	1	1	1	0

Table 3.8: Results of the two sample T-test performed between SROCC values obtained by different measures (Evaluated on CSIQ).

	PSNR	SSIM	NIQE	QAC	BLISS-S	BLISS-C
SROCC	0.0328	0.0175	0.0180	0.0237	0.0164	0.0143
LCC	0.0302	0.0181	0.0160	0.0243	0.0152	0.0137

Table 3.9: Standard deviation of SROCC and LCC for 1000 iterations of experiments on LIVE.

CORNIA [19] respectively. We train these models on images with JP2k, JPEG, WN and GBLUR distortions in the LIVE dataset and test on the images with the same four types of distortions in the TID2008 dataset and the CSIQ dataset. The SROCC and LCC are evaluated on all four types of distortions from 1000-fold cross-validation experiments and median values are reported in Tables 3.10 and 3.11. In this experiment, CORNIA is trained using a linear SVR with the parameters specified in Table 3.2 and BRISQUE is trained using SVR with a RBF kernel<sup>1</sup>. As is shown in Tables 3.10 and 3.11, models trained on the synthetic scores are comparable to models trained on the human opinion score. BLISS works well with both CORNIA and BRISQUE. The best performance is achieved by training on the synthetic scores. This result implies that we can replace human opinion scores with

---

<sup>1</sup>Parameters for training BRISQUE model using libsvm are suggested by the author as “-b 1 -s 3 -g 0.05 -c 1024 -p 1”.

SROCC	CORNIA	BRISQUE	LCC	CORNIA	BRISQUE
DMOS	0.881	0.882	DMOS	0.883	0.892
SS	<b>0.905</b>	0.897	SS	<b>0.925</b>	0.893

Table 3.10: Train on LIVE and test on TID2008

SROCC	CORNIA	BRISQUE	LCC	CORNIA	BRISQUE
DMOS	0.899	0.899	DMOS	0.914	0.927
SS	<b>0.908</b>	0.895	SS	<b>0.928</b>	0.912

Table 3.11: Train on LIVE and test on CSIQ

synthetic scores without loss of performance.

### 3.4.3.3 Comparison of the combined synthetic score and FR measures

#### Evaluation on LIVE

To demonstrate the effectiveness of our score combination method, we test the five FR measures and the combined measures on the LIVE dataset [36]. Table 3.12 shows the LCC and SROCC obtained using each FR measure independently and the synthetic score based on the corresponding FR measures<sup>2</sup>. As shown in this table, by exploiting the overall rank information, combined measures consistently improve over each individual measure. All five FR measures have high SROCC on LIVE, but the combined measures slightly outperforms their base measures. The LCC values are also significantly improved. The conventional method for improving the LCC of a FR measure relies on fitting a nonlinear logistic function, but human

---

<sup>2</sup>No nonlinear fitting procedure is applied in this experiment.

SROCC	GMSD	VIF	FSIM	FSIMC	WSSIM
original	0.960	0.964	0.963	0.965	0.957
SS, $\lambda_0 = 1$	0.967	0.970	0.968	0.968	0.966
SS, $\lambda_0 = 4$	0.969	0.970	0.969	0.968	0.968
LCC	GMSD	VIF	FSIM	FSIMC	WSSIM
original	0.942	0.941	0.859	0.860	0.803
SS, $\lambda_0 = 1$	0.965	0.958	0.956	0.956	0.945
SS, $\lambda_0 = 4$	0.967	0.963	0.967	0.967	0.966

Table 3.12: Test FR measures on LIVE (779 distorted images): ‘original’–correlation between original FR measures and DMOS; ‘SS’–correlation between synthetic scores and DMOS.

opinion scores are required to find the optimal parameters in the logistics function. Our method improves LCC in a fully unsupervised way. One key factor to the success of BLISS is that BLISS use synthetic scores that have high linear correlation with human opinion score to train the BIQA model.

Next we examine the effect of the balancing constant  $\lambda_0$ . Figs.3.2 and 3.3 show how the SROCC and LCC of the combined scores change with different values of  $\lambda_0$ .  $FR$ -SS represents the synthetic score with  $FR$  as the base measure.  $\lambda_0 = 0$  corresponds to using original FR measures. When  $\lambda_0$  is very small, the synthetic score is dominant by the base FR measure and the performance is primarily determined by the base FR measure. As we increase the value of  $\lambda_0$ , the importance of the rank information increases. When  $\lambda_0 \geq 1$ , the value of LCC and SROCC is not very sensitive to the value of  $\lambda_0$ .

To demonstrate that our method is robust to the dataset size, we randomly sample a subset of the LIVE dataset and apply our method on the subset. This process is repeated 1000 times and median values of SROCC and LCC are presented

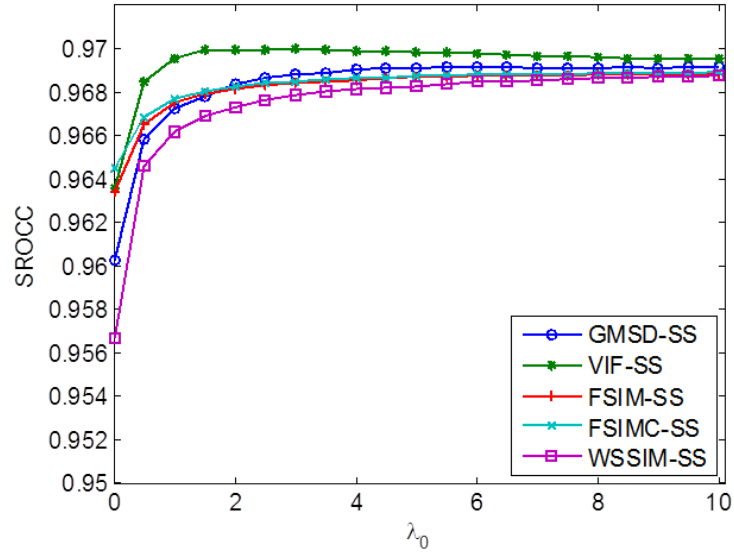


Figure 3.2: Effect of  $\lambda_0$  on SROCC (Tested on LIVE).

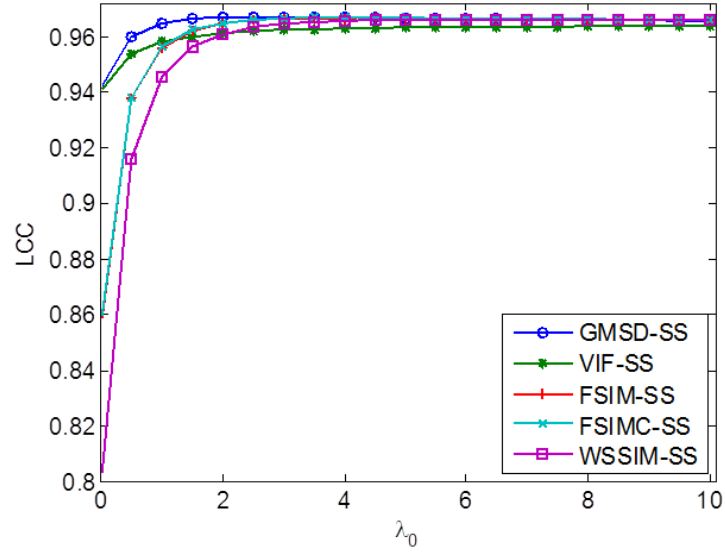


Figure 3.3: Effect of  $\lambda_0$  on LCC (Tested on LIVE).

in Figs. 3.4 and 3.5. We can see that the performance decreases only slightly as we reduce the dataset size to 10% of original size.

### Evaluation on TID2008



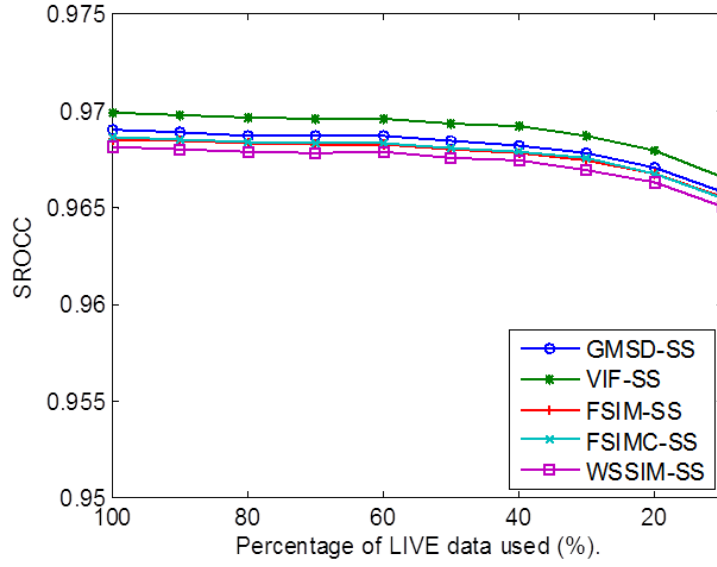


Figure 3.4: Effect of dataset size on SROCC (Tested on LIVE,  $\lambda_0 = 4$ ).

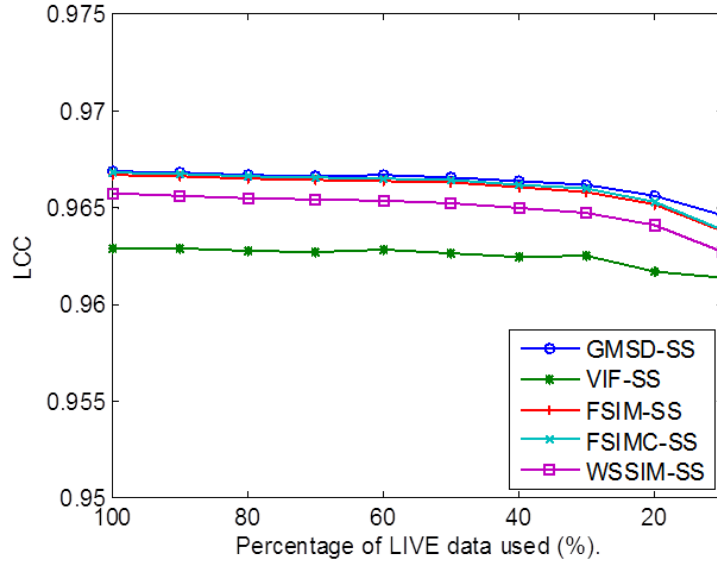


Figure 3.5: Effect of dataset size on LCC (Tested on LIVE,  $\lambda_0 = 4$ ).

All the five FR measures have similar performance in terms of SROCC on the LIVE dataset. However, on the TID2008 dataset, the performance of the five FR measures varies a lot. We compute the pair-wise SROCC between the FR measures and look

SROCC	GMSD	FSIM	FSIMC	WSSIM
original	0.891	0.881	0.884	0.856
SS, $\lambda_0 = 1$	0.898	0.892	0.892	0.887
SS, $\lambda_0 = 4$	0.896	0.893	0.893	0.892
LCC	GMSD	FSIM	FSIMC	WSSIM
original	0.872	0.830	0.834	0.809
SS, $\lambda_0 = 1$	0.885	0.884	0.884	0.884
SS, $\lambda_0 = 4$	0.878	0.878	0.878	0.878

Table 3.13: Test FR measures on TID2008 (1700 distorted images): ‘original’–correlation between original FR measures and DMOS; ‘SS’–correlation between synthetic scores and DMOS.

at the average value of the SROCC. For GMSD, VIF, FISM, FSIMC and WSSIM, the average SROCCs are 0.891, 0.784, 0.930, 0.930, 0.900 respectively. It is obvious that VIF is not consistent with the other four types of FR measures. Therefore, VIF is discarded for computing the RRF score. Table 3.13 presents the evaluation results on 1700 distorted images in the TID2008 dataset. We see that GMSD performs the best among all FR measures and the synthetic scores slightly outperform GMSD.

#### 3.4.3.4 Combining SSIM and WMSE

In order to train an accurate NR-IQA model, we have selected five state-of-the-art FR measures to participate in the score combination method. Next, we show that our method can also significantly boost the performance of base FR measures when the measures anticipated in the combination are relatively “weak” and when only two measures are used in the combinations. In particular, we test on **SSIM** [55] and **WMSE** [54]. Tables 3.14 and 3.15 shows the SROCC and LCC values evaluated on LIVE and TID2008.

Since FR measures are usually not in the same range as the human opinion scores, the LCC between FR measures and human opinion scores is usually low. The conventional way for mapping FR measures to the range of human opinion scores and improving the LCC is to apply a nonlinear fitting procedure. A logistic fitting function,  $Q_p = \beta_1(\frac{1}{2} - \frac{1}{\exp(\beta_2(Q - \beta_3))}) + \beta_4 Q + \beta_5$  is often used for the nonlinear mapping. It requires human opinion scores to find the optimal parameters in the function. Table 3.14 also presents the result obtained by applying the nonlinear fitting on the original FR measure. Optimal parameters of the logistic fitting function for LIVE and TID2008 are trained on corresponding datasets. It can be seen that our unsupervised method outperforms the supervised nonlinear fitting method.

SROCC	SSIM	WMSE	LCC	SSIM	WMSE
original	0.910	0.933	original	0.825	0.554
logistic	0.910	0.933	logistic	0.904	0.714
SS, $\lambda_0 = 1$	0.933	<b>0.950</b>	SS, $\lambda_0 = 1$	0.911	0.924
SS, $\lambda_0 = 4$	<b>0.942</b>	0.948	SS, $\lambda_0 = 4$	<b>0.933</b>	<b>0.943</b>

Table 3.14: Test FR measures on LIVE (779 distorted images): ‘original’–original FR measures and DMOS; ‘logistic’–FR measures with nonlinear fitting; ‘SS’–combined synthetic scores.

SROCC	SSIM	WMSE	LCC	SSIM	WMSE
original	0.775	0.682	original	0.740	0.496
logistic	0.775	0.682	logistic	<b>0.773</b>	0.664
SS, $\lambda_0 = 1$	0.819	0.812	SS, $\lambda_0 = 1$	<b>0.817</b>	0.798
SS, $\lambda_0 = 4$	<b>0.821</b>	<b>0.819</b>	SS, $\lambda_0 = 4$	0.812	<b>0.809</b>

Table 3.15: Test FR measures on TID2008 (1700 distorted images): ‘original’–original FR measures and DMOS; ‘logistic’–FR measures with nonlinear fitting; ‘SS’–combined synthetic scores.

Fig. 3.6 shows scatter plots of (a) MOS vs. SSIM (b) MOS vs. SSIM-SS

(Synthetic score with SSIM as base measure.) and (c) SSIM vs SSIM-SS obtained from TID2008. We can see that by incorporating rank information, the linearity of the FR measure with the human opinion scores is significantly improved.

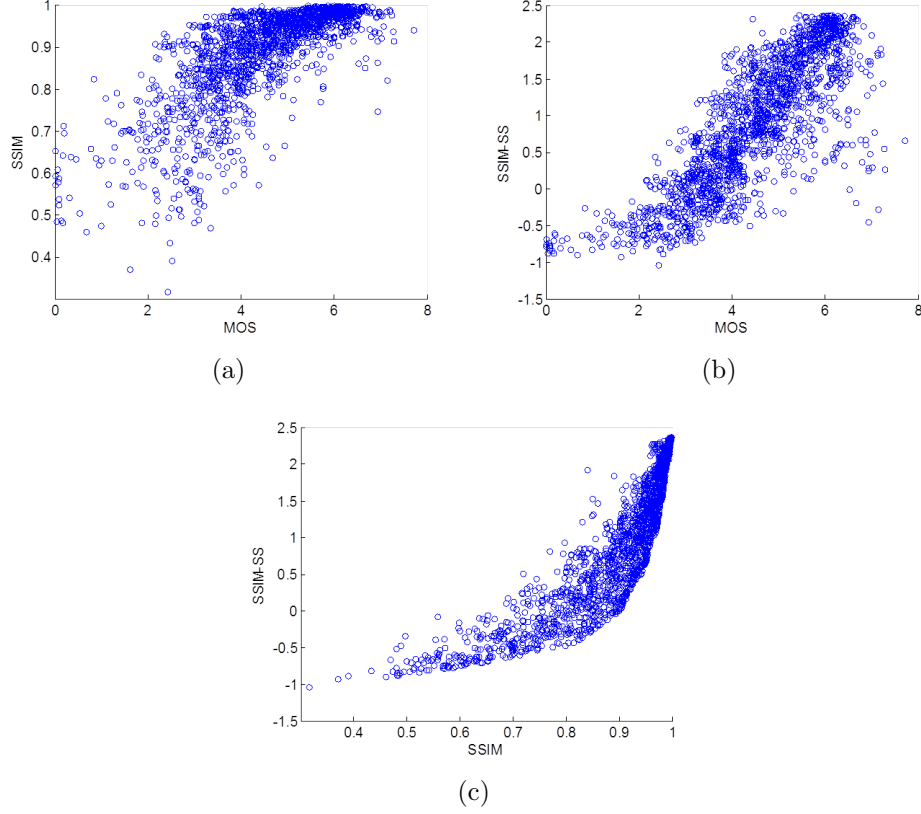


Figure 3.6: TID2008 test: (a) MOS vs. SSIM (b) MOS vs. SSIM-SS (Synthetic score with SSIM as base measure.) (c) SSIM vs SSIM-SS.

## 3.5 Discussion and Extension

### 3.5.1 Discussion

As shown in our experimental results in Section 3.4.3.2, often times models trained on the synthetic scores can outperform models trained on human opinion scores. This result may be explained by the inherent ambiguity of human opinion

scores. The mean opinion score (MOS) test is the most widely used subjective test for obtaining groundtruth data for image quality datasets. However, there are many known problems with the MOS test [61]. In a MOS test, subjects are asked to rate image quality using an ordinal scale: “Bad”, “Poor”, “Fair”, “Good” and “Excellent”, then a numeric score 1–5 is associated with each ordinal label. The average of these numeric scores is taken as the final MOS score for the image. This procedure assumes that the scale is uniform. However this assumption is not true in practice because the cognitive differences between the consecutive MOS scales may not be the same. For example, the difference between “Fair” and “Poor” may not equal to the difference between “Excellent” and “Good”. Furthermore, the MOS rating procedure is somewhat obscure for experimental subjects in that subjects can be easily confused about which score they should give in each test and the resulting absolute judgments can be very noisy. Recently, pair-wise comparison based tests have been proposed as an alternative to the MOS test [61]. How to properly design the subjective IQA test is still an open problem. We may consider the current ground-truth labels in the IQA dataset as a noisy approximation to the unknown “gold-standard”.

The human opinion scores in different datasets were also obtained under different experimental conditions. Therefore, the MOS labels in different datasets may not be consistent. On the other hand, FR measures are objective measures that capture the inherent properties of image distortion which do not vary from dataset to dataset. It is therefore possible to train a better prediction model using synthetic scores.

### 3.5.2 Extension

One extension to our work is to consider how to incrementally update the FR measures given a new image. Given  $N$  images in an IQA dataset, we have described a method to collectively improve their FR measures. Once we have the RRF rank and FR measures of an IQA dataset, given a new image  $I$ , we can adjust its FR measure value in the following way:

1. Compute  $RRFscore(I)$ : this involves finding the ranks of  $I$  in  $K$  sorted list of size  $N$ . We may use binary search with complexity  $O(\log(N))$  to find the rank of  $I$  in one sorted list, therefore the overall complexity is  $O(K\log(N))$ .

2. Given  $RRFscore(I)$ , find its rank in the sorted list of  $RRF$  scores of the  $N$  images in the dataset. The complexity is again  $O(\log(N))$ . Suppose the rank of  $RRFscore(I)$  is  $k$ , we can adjust the FR measures of  $I$  as follows:

$$FR(I)_{new} = FR(I) - \lambda((k - 1) - (N + 1 - k)) = FR(I) - \lambda(2k - N - 2)$$

## Chapter 4: Active Learning for Subjective Image Quality Assessment

### 4.1 Introduction

Estimating gold-standard labels (strengths, scores, etc.) based on subjective judgments provided by humans is a critical step in psychological experiments with applications in many research fields [62]. We study the problem of Quality of Experience (QoE) evaluation, which in general aims to obtain subjective satisfaction of user’s experience with a service (for example, web browsing, phone calls, video chatting or online shopping.) or with some multimedia content (for example, videos, images, etc.). In particular, we investigate image quality assessment (IQA) problem, but our method can be applied to any general problem of QoE evaluation.

Absolute Category Rating [63] is one of the most popular subjective IQA tests. It consists of having a panel of subjects rate images using an ordinal scale: 1-“Bad”, 2-“Poor”, 3-“Fair”, 4-“Good” and 5-“Excellent”. For a given image, its score is computed as the average scores from all subjects. This is also known as the Mean Opinion Score (MOS). Despite the popularity of the MOS test, there are many known problems [61, 64]. First, most previous work in QoE [65] treats the MOS

scale as an interval scale instead of ordinal scale and assumes that the cognitive distances between the consecutive MOS scales are the same. However, assumptions such as: “Fair”-“Poor”=“Good”-“Fair”, are not always true in practice. Second, absolute rating procedures are somewhat obscure so subjects can be easily confused about which scale they should give in each test and different subjects may have different interpretations of the scale. Therefore the resulting rating observations can be very noisy.

To overcome the limitation of the MOS test, the Paired Comparison (PC) test [61,62,64,66–69] has been proposed as an alternative. In the simplest configuration, two images  $A$  and  $B$  are shown to a subject who is asked to “prefer” one of them. Compared to the rating test, making a decision in a paired comparison test is much simpler and less confusing for the subject. However, when  $n$  images need to be compared, the total number of pairs is  $\binom{n}{2}$  and when  $n$  is large, the cost for obtaining a full set of pairwise comparisons is prohibitively expensive. HodgeRank on Random Graphs (HRRG) [68,69] has been introduced to reduce the cost of the PC test by using a random sampling method with HodgeRank [26]. We approach this problem differently by combining the MOS test and the PC test via active sampling. As will be shown experimentally, our method significantly outperforms the HRRG for crowdsourcing subjective IQA.

Our method is motivated by the following observations: 1) Although the MOS test may not be able to accurately rank two images with similar quality due to the observation noise, it can provide an estimate of the underlying quality score at a coarse level. 2) In the PC test, we explicitly ask humans to compare pairs of images,



therefore the PC test can provide fine discrimination on images with similar quality.

3) Once we have some coarse estimates on the underlying scores, a complete set of PC test would be unnecessary. For example, it would be unnecessary to perform a paired comparison on an image with MOS score 1 and an image with MOS score 5, since we can already tell the difference with high confidence. Based on these observations, we will show that combining the MOS and PC tests will provide a more efficient design for subjective IQA. In this chapter, we will answer the following two questions:

1. Given a collection of observations from the MOS test and the PC test, how can we combine them to estimate the underlying score?
2. In both laboratory studies and crowdsourced settings, subjective judgments are obtained at a defined cost. How can we effectively sample a subset of MOS and PC tests so that we can achieve desired accuracy with minimal cost?

## 4.2 Related Work

### 4.2.1 Crowdsourceable QoE

Conventional subjective QoE experiments conducted in laboratory settings can be expensive and time-consuming and typically only a small number of subjects are involved. With ubiquitous internet access and the rise of internet micro-labor markets supported by systems such as Amazon Mechanical Turk, there has been an increase in interest in designing subjective QoE tests for crowdsourced settings.

Previous work on Crowdsourceable QoE considers the MOS and PC tests

independently. Ribeiro et al. [65] performed the MOS test for QoE assessment using crowdsourcing. They developed a two-way random effects model to model the uncertainty in subjective tests and proposed a post-screening method and rewarding mechanism to facilitate the process. Chen et al. [61] proposed a crowdsourcable QoE assessment framework for multimedia content, in which interval-scale scores are derived from a full set of paired comparisons. However, since a complete set of paired comparisons has to be performed, this method cannot be applied on a large-scale. To address this problem, Xu et al. [68, 69] introduced the HodgeRank on Random Graphs (HRRG) test, where random sampling methods based on Erdős-Rényi random graphs were used to sample pairs and the HodgeRank [26] was used to recover the underlying quality scores from the incomplete and imbalanced set of paired comparisons. This method can effectively reduce the cost of PC tests required for achieving a certain accuracy. We will show experimentally that by combining information from MOS and PC tests via active sampling, we can further reduce the cost of experiments.

#### 4.2.2 Preference Aggregation

The problem we are trying to solve is essentially an information aggregation problem, where we want to integrate information from multiple sources into a consensus score. The problem of preference aggregation has been extensively studied in the information retrieval community [57, 70–72]. In particular, there has been some recent work in this field that applied active learning for preference aggregation.

Given a pair of objects, a utility function is defined that measures the “usefulness” of performing a paired comparison. Then pairs with high utilities are chosen as queries and sent to an oracle or to human subjects.

Pfeiffer et al. [71] introduced an active learning method based on the Thurstone-Mosteller model [66, 73] for pairwise rank aggregation. At each iteration of an experiment, this method adaptively chooses one pair of objects to compare. The paper shows the advantage of using active sampling over random sampling. Chen et al. [70] proposed an active learning model based on the Bradley-Terry Model [74] which adopts an efficient online Bayesian updating scheme that does not require retraining of the whole model when new observations are obtained. All of these previous works focus solely on aggregating information obtained from PC tests. A single optimal pair is usually chosen at each iteration of the experiment. This is inefficient in a crowdourced setting, where multiple subjects may work in parallel and workers may expect to work on multiple tests instead of taking one single test in each working session. It is desirable to develop a batch-mode active learning method for the crowdsourcable subjective QoE problem.

Gleich and Lim [75] introduced several ad-hoc methods for building a preference matrix from rating observations based on the arithmetic mean of score differences, geometric mean of score ratios, binary comparisons, strict binary comparisons and logarithmic odds ratios. We may apply these methods to convert the rating observations into the preference observations. However, it is not clear how to measure the utility of the MOS test. Our method combines the MOS test and the PC test directly via a unified probabilistic model and the utility of each individual MOS and

PC test is defined as the expected information gain given by the test.

### 4.3 Combining Ratings and Paired Comparisons

This section presents the probabilistic model for combining the MOS test and the PC test. Suppose we have  $n$  images  $A_1, A_2, \dots, A_n$  with underlying scores  $s = (s_1, s_2, \dots, s_n)$ . We model a subject's perceived quality of image  $A_i$  as a random variable:  $r_i = s_i + \varepsilon_i$ , where the noise term is a Gaussian random variable  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ . In the remainder of this section, we first derive the likelihood functions of the underlying score given the MOS and PC observations independently. We then present a hybrid system which estimates the underlying score using Maximum A Posteriori Estimation (MAP).

#### 4.3.1 Mean Opinion Score Test

Thurstone's law of categorical judgment [62] is applied for analyzing the rating observations. Assume the perceived categorical observation for  $A_i$  is  $m_i$  and  $m_i \in \mathcal{M}$ , where  $\mathcal{M}$  is a finite set of  $K$  ordered categories and  $K = 5$  in the case of the MOS test. Without loss of generality, these categories are denoted as consecutive integers:  $\mathcal{M} = \{1, 2, \dots, K\}$ . We further introduce a set of cutoff values  $-\infty \equiv \gamma_0 < \gamma_1 < \gamma_2 < \dots < \gamma_{K-1} < \gamma_K \equiv \infty$ <sup>1</sup>. When  $r_i$  falls between the cutoffs  $\gamma_{c-1}$  and  $\gamma_c$ , the observed categorical label is  $c$ , i.e.  $m_i = c$ , we have

---

<sup>1</sup>The original law of categorical judgment [62] assumes the randomness of  $\gamma_c$ , for simplicity, we assumes  $\gamma_c$  to be deterministic as in [76].

$$\begin{aligned}
Pr(m_i|s_i) &= Pr(\gamma_{m_i-1} < s_i + \varepsilon_i \leq \gamma_{m_i}) \\
&= \Phi\left(\frac{\gamma_{m_i}-s_i}{\sigma}\right) - \Phi\left(\frac{\gamma_{m_i-1}-s_i}{\sigma}\right)
\end{aligned} \tag{4.1}$$

where  $\Phi(\cdot)$  represents Cumulative Density Function (CDF) of standard Gaussian distribution.

In the MOS test, repeated observations are made for each image. We define the rating observation matrix  $M$  as follows:

$$M = \begin{pmatrix} M_{1,1} & M_{2,1} & \cdots & M_{n,1} \\ M_{1,2} & M_{2,2} & \cdots & M_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ M_{1,K} & M_{2,K} & \cdots & M_{n,K} \end{pmatrix} \tag{4.2}$$

where  $M_{i,j}$  is the number of times the image  $A_i$  is observed as in the  $j$ -th category. Given the underlying score  $s$ , we assume the categorical observations of each image are conditionally independent. We then have the probability of observing  $M$  as follows:

$$\begin{aligned}
Pr(M|s) &= \prod_{i=1}^n Pr(M_{i,1}, M_{i,2}, \dots, M_{i,K}|s_i) \\
&= \prod_{i=1}^n \binom{M_{i,1} + \dots + M_{i,K}}{M_{i,1}, \dots, M_{i,K}} \prod_{k=1}^K Pr(m_i = k|s_i)^{M_{i,k}} \\
&= c_1 \prod_{i=1}^n \prod_{k=1}^K \left( \Phi\left(\frac{\gamma_k - s_i}{\sigma}\right) - \Phi\left(\frac{\gamma_{k-1} - s_i}{\sigma}\right) \right)^{M_{i,k}}
\end{aligned} \tag{4.3}$$

where  $c_1$  is a constant.

### 4.3.2 Paired Comparison Test

In the PC test, if the perceived score  $r_i > r_j$ , we say that  $A_i$  is preferred to  $A_j$ , which is denoted as  $A_i \succ A_j$ . The probability of  $A_i \succ A_j$  is given by:

$$Pr(A_i \succ A_j) = Pr(s_i + \varepsilon_i > s_j + \varepsilon_j) = \Phi\left(\frac{s_i - s_j}{\sqrt{2}\sigma}\right) \quad (4.4)$$

Eq. 4.4 is known as the Thurstone-Mosteller Case V model [66,73]. Preferences obtained from a set of PC tests can be characterized by a preference matrix and we define the preference matrix  $P$  as:

$$P = \begin{pmatrix} 0 & P_{1,2} & \cdots & P_{1,n} \\ P_{2,1} & 0 & \cdots & P_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n,1} & P_{n,2} & \cdots & 0 \end{pmatrix} \quad (4.5)$$

where  $P_{i,j}$  is the number of times  $A_i \succ A_j$  is observed. Then the probability of observing  $P$  is:

$$\begin{aligned}
Pr(P|s) &= \prod_{i,j \in 1, \dots, n, i < j} Pr(P_{i,j}, P_{j,i} | s_i, s_j) \\
&= \prod_{i < j} \binom{P_{i,j} + P_{j,i}}{P_{i,j}} Pr(A_i \succ A_j)^{P_{i,j}} Pr(A_j \succ A_i)^{P_{j,i}} \\
&= \prod_{i < j} \binom{P_{i,j} + P_{j,i}}{P_{i,j}} \Phi\left(\frac{s_i - s_j}{\sqrt{2}\sigma}\right)^{P_{i,j}} \Phi\left(\frac{s_j - s_i}{\sqrt{2}\sigma}\right)^{P_{j,i}} \\
&= c_2 \prod_{i \neq j} \Phi\left(\frac{s_i - s_j}{\sqrt{2}\sigma}\right)^{P_{i,j}}
\end{aligned} \tag{4.6}$$

where  $c_2$  is a constant.

### 4.3.3 Posterior Probability of the Underlying Score

Given observations from both MOS and PC tests, the hybrid system estimates the underlying score by maximizing the posterior probability

$$\hat{s} = \operatorname{argmax}_s Pr(s|P, M) \tag{4.7}$$

Computing  $Pr(s|P, M)$  is not a trivial task. The likelihood functions in Eq. 4.3 and Eq. 4.6 are conditioned on several unknown model parameters including: noise variance  $\sigma$  and cut-off parameters  $\gamma_1, \dots, \gamma_{K-1}$ . Since the likelihood functions are scale-invariant, i.e.  $Pr(M|s, \gamma, \sigma) = Pr(M|ks, k\gamma, k\sigma)$  and  $Pr(P|s, \sigma) = Pr(P|ks, k\sigma)$  for a constant  $k \neq 0$ , without loss of generality, we may fix  $\sigma = 1/\sqrt{2}$ . With  $\sigma$  fixed, the likelihood functions are still translation-invariant, i.e.  $Pr(M|s, \gamma) = Pr(M|s + k, \gamma + k)$  and  $Pr(P|s) = Pr(P|s + k)$  for a constant  $k$ . To

make the objective identifiable, we further assume  $\gamma_1 = 0$ .  $K - 2$  model parameters  $\gamma_2, \dots, \gamma_{K-1}$  remain unknown. We denote the set of unknown model parameters  $\gamma = \{\gamma_2, \dots, \gamma_{K-1}\}$ .

In a full Bayesian treatment, computing  $Pr(s|P, M)$  requires integrating the model parameters over all possible values, which can be implemented using Monte Carlo methods. However, these computations might be prohibitively expensive. Alternatively, we approximate  $Pr(s|P, M)$  by  $Pr(s|P, M, \hat{\gamma})$  where  $\hat{\gamma}$  refers to the optimal setting of  $\gamma$ . Specifically,  $\hat{\gamma} = \text{argmax}_{\gamma} Pr(M, P|\gamma)$ , which is the Maximum Likelihood Estimate of  $\gamma$ . To obtain an analytical form of the gradients of  $Pr(M, P|\gamma)$  w.r.t  $\gamma$  and a Gaussian form approximation to the posterior probability  $Pr(s|M, P, \hat{\gamma})$ , we apply the Laplace approximation [77]. To illustrate the approximation procedure, let us define:

$$\mathcal{F}_{\gamma}(s) = -\log Pr(M|s, \gamma) - \log Pr(P|s) - \log Pr(s) \quad (4.8)$$

where we assume a Gaussian prior on  $s \sim N(\mu, \Omega)$ . The Hessian matrix of  $\mathcal{F}_{\gamma}(s)$  is given by:

$$R_{\gamma}(s) = \frac{\partial^2 \mathcal{F}_{\gamma}(s)}{\partial s \partial s^T} \quad (4.9)$$

Denoting the minimizer of  $\mathcal{F}_{\gamma}(s)$  as  $\hat{s}_{\gamma}$  and  $\hat{R}_{\gamma} = R_{\gamma}(\hat{s}_{\gamma})$ , applying a Laplace approximation, we have



$$\mathcal{F}_\gamma(s) \approx \mathcal{F}_\gamma(\hat{s}_\gamma) + \frac{1}{2}(s - \hat{s}_\gamma)^T \hat{R}_\gamma (s - \hat{s}_\gamma) \quad (4.10)$$

Using the above approximation,  $Pr(M, P|\gamma)$  can be computed analytically as follows:

$$Pr(M, P|\gamma) = \int Pr(s) Pr(M|s, \gamma) Pr(P|s) ds \quad (4.11)$$

$$= \int \exp(-\mathcal{F}_\gamma(s)) ds \approx \exp(-\mathcal{F}_\gamma(\hat{s}_\gamma)) (2\pi)^{\frac{n}{2}} |\hat{R}_\gamma|^{-1/2}$$

Using Eq. 4.11, the gradients of the  $\log(Pr(M, P|\gamma))$  w.r.t  $\gamma$  can be computed analytically. Details of the gradient computation is presented in Appendix A. Gradient-based optimization methods can be used to find MLE of  $\gamma$ .

Given the optimal cut-off parameter  $\hat{\gamma}$ , the posterior probability of  $s$  can be approximated by:

$$Pr(s|P, M) \propto Pr(M|s, \hat{\gamma}) Pr(P|s) Pr(s) \quad (4.12)$$

$$= \exp(-\mathcal{F}_{\hat{\gamma}}(s)) \propto N(\hat{s}_{\hat{\gamma}}, \hat{R}_{\hat{\gamma}}^{-1})$$

The MAP estimate of  $s$  is  $\hat{s}_{\hat{\gamma}}$ . In order to ensure a global optimal solution of the MAP estimate, Eq. 4.8 has to be a convex function. It has been shown in [76] that  $-\log Pr(M|s, \gamma) - \log(Pr(s))$  is convex. However, in order to make sure  $-\log(Pr(P|s))$  has a unique minimizer, Ford's condition [78] has to be satisfied. In practice, this can be achieved by adding a small constant to each zero-valued element in the preference matrix  $P$ . This is also known as smoothing.

## 4.4 Subjective Experimental Design based on Active Sampling

Subjective judgments are usually obtained at a certain cost and it is therefore desirable to design cost-efficient experiments. We introduce an active random sampling method which constructs a set of queries consisting of MOS and PC tests based on the expected information gain provided by each. Let  $\mathcal{E}_i$  denote the experiment which makes one absolute judgment on the object  $A_j$  and  $\mathcal{E}_{ij}$  be the experiment that makes a pairwise comparison between  $A_i$  and  $A_j$ .

### 4.4.1 Information Measure of Experiments

The purpose of experiments is to gain knowledge about the state of nature. We adopt the Bayesian Optimal Design framework introduced by Lindley [79] and evaluate an experiment using the Expected Information Gain (EIG) provided by conducting this particular experiment. In the subjective IQA problem, the state of nature (or parameter) to be estimated is the quality score  $s = \{s_1, \dots, s_n\}$ . Before conducting the experiment  $\mathcal{E}$ , our knowledge of  $s$  is characterized by the prior distribution of  $s \sim Pr(s)$ . The EIG provided by an experiment  $\mathcal{E}$  is denoted  $I(\mathcal{E}, Pr(s))$ . The general formula of  $I(\mathcal{E}, Pr(s))$  is given by [79]:

$$I(\mathcal{E}, Pr(s)) = E_s \left[ \int \log \left\{ \frac{Pr(x|s)}{Pr(x)} \right\} Pr(x|s) dx \right] \quad (4.13)$$

where  $E_s(\cdot)$  is expectation taken w.r.t  $Pr(s)$ . For the MOS test, suppose the outcome of  $\mathcal{E}_i$  is  $x_i \in \{1, 2, \dots, K\}$  and  $p_{ik} = P(x_i = k|s)$ . It is easy to verify that

$p(x_i = k) = E_s(p(x_i = k|s)) = E_s(p_{ik})$  and we have

$$I(\mathcal{E}_i, Pr(s)) = E_s[\sum_{k=1}^K p_{ik} \log(\frac{p_{ik}}{p(x_i=k)})] \quad (4.14)$$

$$= E_s[\sum_{k=1}^K p_{ik} \log(p_{ik})] - \sum_{k=1}^K E_s(p_{ik}) \log E_s(p_{ik})$$

For the PC test, suppose the outcome of  $\mathcal{E}_{ij}$  is  $x_{ij}$  and  $x_{ij} = 1$  if  $A_i \succ A_j$ ;  $x_{ij} = 0$  if  $A_i \prec A_j$ . Define  $p_{ij} = p(x_{ij} = 1|s)$  and  $q_{ij} = 1 - p_{ij}$ . It is easy to verify that  $p(x_{ij} = 1) = E_s(p(x_{ij} = 1|s)) = E_s(p_{ij})$  and  $p(x_{ij} = 0) = E_s(q_{ij})$ . The information gain provided by  $\mathcal{E}_{ij}$  is:

$$\begin{aligned} I(\mathcal{E}_{ij}, Pr(s)) &= E_s[p_{ij} \log(\frac{p_{ij}}{p(x_{ij}=1)}) + q_{ij} \log(\frac{q_{ij}}{p(x_{ij}=0)})] \\ &= E_s[p_{ij} \log(p_{ij}) + q_{ij} \log(q_{ij})] \end{aligned} \quad (4.15)$$

$$-E_s(p_{ij}) \log(E_s(p_{ij})) - E_s(q_{ij}) \log(E_s(q_{ij}))$$

Assuming that judgments on a pair of images are conditionally independent given the underlying score, it has been proven in [80] that the information gain obtained by a set of paired comparisons is the sum of contributions from each pair. It is worth noting that the prior distribution  $Pr(s)$  is actually conditioned on previous observations as in Eq. 4.12, but we omit the conditions here for ease of representation. In Eq. 4.12, we introduced a Gaussian approximation to the posterior distribution. Therefore, we can use the Gauss-Hermite quadrature [81] to compute the expectation efficiently. Fig. 4.1 shows how the EIG  $I(\mathcal{E}_{ij}, Pr(s))$  changes with

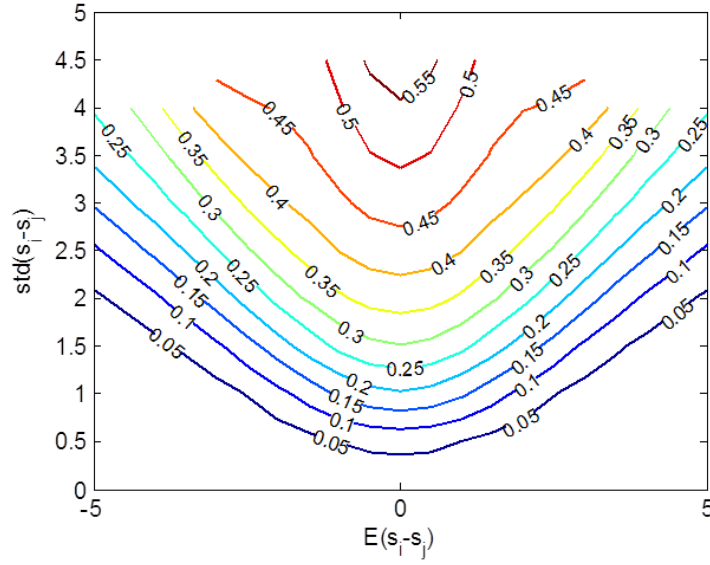


Figure 4.1: Contour plot of  $I(\mathcal{E}_{ij}, Pr(s_i, s_j))$  as function of the expectation and the standard deviation of  $s_i - s_j$ .

the expectation and the standard deviation of  $s_i - s_j$ . We can see that the utility of the PC test increases as  $E(s_i - s_j)$  decreases and  $std(s_i - s_j)$  increases. This implies that the EIG obtained by performing a PC test on two images with similar quality is higher than that for those with very different quality.

#### 4.4.2 Active Sampling

Suppose we have  $n$  images, we would like to make one observation related to each image at each round of the experiment. For an image  $A_i$ ,  $n$  different tests related to  $A_i$  can be conducted including one MOS test  $\mathcal{E}_i$  and  $n - 1$  PC tests  $\mathcal{E}_{ij}, j = 1, \dots, n, j \neq i$ . We select the test which has the highest EIG to perform. At the  $t$ -th iteration of the experiment, the test selected to perform for image  $A_i$  is:

$$\mathcal{E}^t(A_i) = \operatorname{argmax}_{\mathcal{E} \in \{\mathcal{E}_{ij}, \mathcal{E}_i | j \neq i\}} I(\mathcal{E}, Pr(s|M_{t-1}, P_{t-1})) \quad (4.16)$$

where  $M_{t-1}$  and  $P_{t-1}$  summarize all rating and preference observations in the previous iterations of the experiment.

Our model assumes that the observation noise has the same standard deviation  $\sigma = 1/\sqrt{2}$  for both the MOS test and the PC test. However, in practice, this assumption may not be true and the MOS test is usually associated with higher noise level. Therefore the utility of performing a MOS test computed under this assumption may be higher than its true utility. A quick fix can be applied by imposing a slightly higher  $\sigma$  when computing the EIG for the MOS test. In particular, we set  $\sigma_{mos} = \frac{1}{\sqrt{2}}(1 + \alpha)$ , where  $\alpha$  is a small positive constant.

## 4.5 Experiments

In this section, we evaluate our method for the crowdsourcing subjective IQA task.

### 4.5.1 Dataset

We have not found any publicly available dataset with a large set of rating and preference judgments for crowdsourcing QoE problems, so we built our own dataset starting with the LIVE IQA dataset [37] for experiments. The LIVE IQA dataset includes 779 distorted images with five different types of distortions derived from 29 reference images. For this work, we selected a subset of 120 images from

the Fast-Fading category. The 120 images include 20 undistorted reference images and 100 distorted images derived from the 20 reference images. Each image in the LIVE dataset is associated with a subjective DMOS score which was obtained through the MOS test. Note that we will not use the DMOS as groundtruth for our experiments, since the accuracy of the DMOS is limited by the nature of the MOS test. Alternatively, we will generate more realistic groundtruth through our new experimental design.

The subjective judgments of the set of images were obtained using the Amazon Mechanical Turk (MTurk) platform<sup>2</sup>. In the MOS test, images are labeled by five ordinal scales: “Bad”, “Poor”, “Fair”, “Good” and “Excellent”. For each image, we collected 50 rating observations and a total of 6000 rating scores for 120 images were obtained from 86 subjects. A complete set of paired comparisons of this dataset includes  $\binom{120}{2} = 7140$  pairs. For each pair, we collected five repeated observations for a total of 35700 pairs from 196 subjects. Examples of the test interface for the MOS test and the PC test are shown in Figs 4.2 and 4.3 respectively.

Using MTurk, each working session is considered one HIT (human intelligence task). In our studies, each HIT includes 10 images for the MOS test or 10 pairs for the PC test. Images were randomly permuted to display in each HIT and for the PC test, the display order of a pair of images was also randomized. Additionally, the maximal number of HITs of PC tests that could be done by one worker was limited to 40 so that we would not have a large set of paired comparisons from the same subject. We will make the dataset publicly available upon publication of our

---

<sup>2</sup><https://www.mturk.com/>

Score	Quality	Deterioration
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable




Image	Score
1	Unrated
2	Unrated
3	Unrated
4	Unrated
5	Unrated
6	Unrated
7	Unrated
8	Unrated
9	Unrated
10	Unrated

ACCEPT the HIT to enable

Want to work on this HIT?  Want to see other HITs?

Figure 4.2: MOS test in the Crowdsourcing experiment.

technical article on the subject.

## 4.5.2 Evaluation Measure

We use the Kendall's  $\tau$  Coefficient and Linear Correlation Coefficient (LCC) for evaluating the performance of subjective tests. Given two global scores on a set

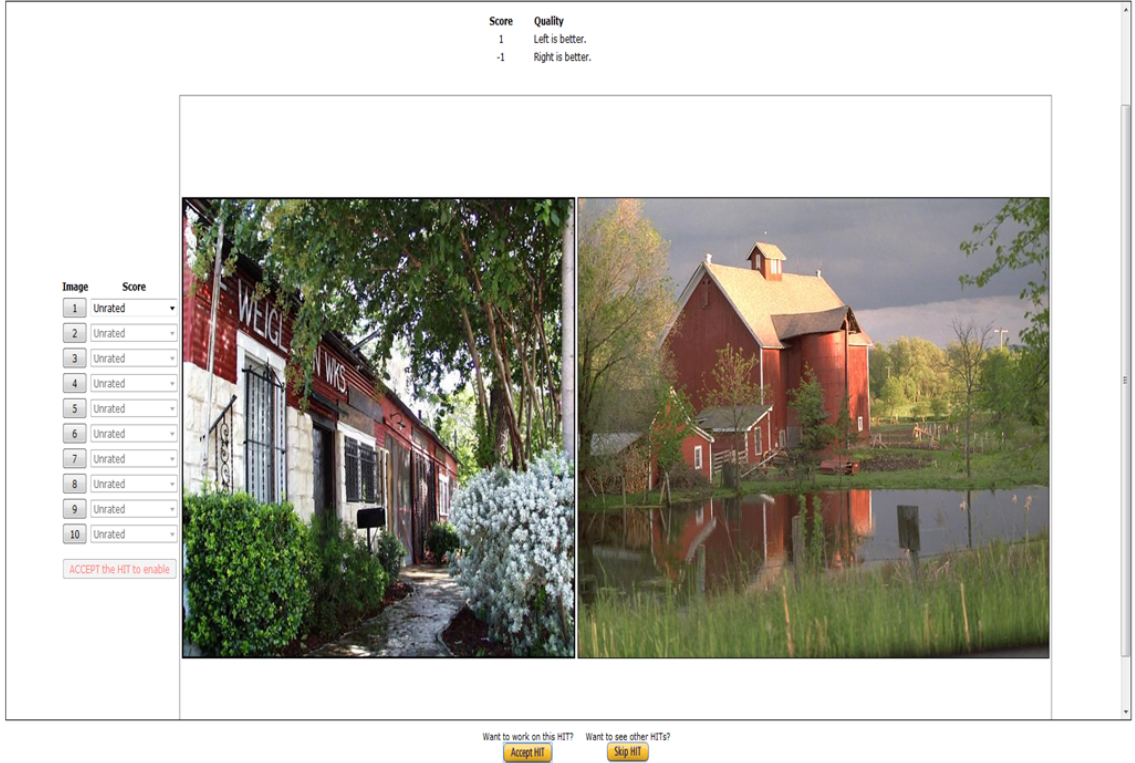


Figure 4.3: PC test in the Crowdsourcing experiment.

of images  $x_i$  and  $y_i$   $i = 1, \dots, n$ , Kendall's  $\tau$  coefficient is defined as

$$\tau(x, y) = \frac{\sum_{i \neq j} X_{ij} Y_{ij}}{\frac{1}{2}n(n-1)} \quad (4.17)$$

where  $X_{ij} = \text{sign}(x_i - x_j)$  and  $Y_{ij} = \text{sign}(y_i - y_j)$ . A pair is concordant if  $X_{ij} = Y_{ij}$  and is discordant otherwise.  $\tau(x, y)$  measures the percentage of concordance pairs minus the percentage of discordant pairs. For identical rankings  $\tau(x, x) = 1$  and for reversed rankings  $\tau(x, -x) = -1$ .

LCC estimates the strength of linear relationship between  $x$  and  $y$ . A high value of  $LCC(x, y)$  does not necessarily imply a high  $\tau(x, y)$ . Kendall's  $\tau$  coefficient is a stricter measure in that it is based on pairwise comparisons.



### 4.5.3 GroundTruth

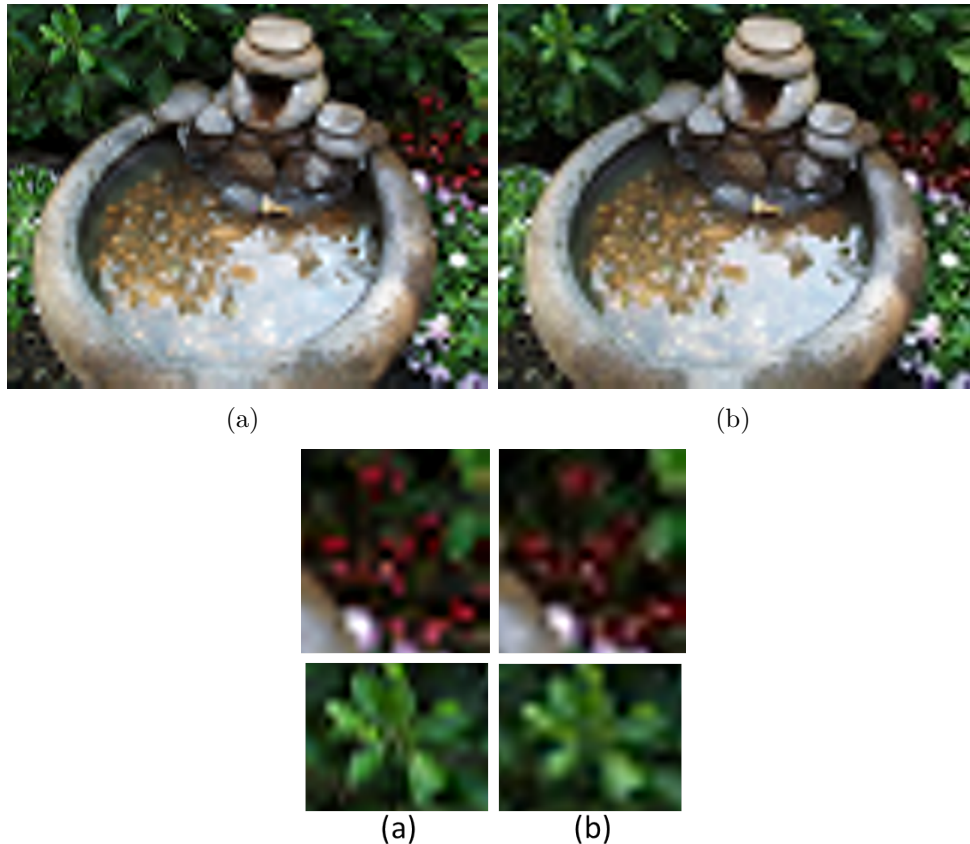


Figure 4.4: Example of Image Pairs.

Groundtruth is obtained by solving the MAP problem described in Section 4.3.3 given all observations, including 6000 rating observations and 35700 preference observations. A smoothing constant 0.5 was added to each zero-valued entry in the preference matrix  $P$  obtained from the PC test. The observation matrices  $M$  and  $P$  are normalized so that  $P(i, j) + P(j, i) = 1$  and  $\sum_{k=1}^5 M(k, i) = 1$ . We use Ipopt [82] for computing the MAP estimate of the underlying score, i.e the minimizer of Fig.4.8. The prior distribution of the underlying score is specified by an uninformative prior defined  $N(\mu, \Omega)$ , where  $\mu = \mathbf{0}$ ,  $\Omega = 1000 \times \mathbf{I}$  and  $\mathbf{I}$  is an

	MOS	PC	HYBRID
Kendall	0.831	0.858	0.859
LCC	0.972	0.978	0.978

Table 4.1: Correlation with LIVE’s DMOS

identity matrix. In addition to the MAP score, we also compute the estimated scores from the MOS test and the PC test independently. For the PC test, HodgeRank [26] is used to estimate the underlying scores. The correlations between the estimated scores obtained from our crowdsourced data and the aligned DMOS<sup>3</sup> provided in the LIVE are shown in Table 4.1. It can be seen that the *LCC* values are high, but the Kendall’s  $\tau$  correlation values are relatively low. Examples of image pairs that DMOS disagrees with our estimates are shown in Fig. 4.4. In this case, the two images are fairly close in quality. In our studies, the first image is preferred to the second image. Taking a closer look at these two images, the first image preserves more detailed information and the second image is more blurred. However, the first image has slightly higher ringing effect compared to the second image. Making a preference judgment between these two images is a highly subjective task and different subjects may have different preferences. In our study, it seems the subjects are more sensitive to blur distortion and prefer sharper images.

#### 4.5.4 Evaluation

To test the performance of our hybrid system with active sampling, we simulate the crowdsourcing experiment by repeatedly and randomly sampling from real judg-

---

<sup>3</sup>By default, DMOS refers to aligned DMOS. The raw DMOS is more noisy.

ments collected from the MTurk. All the raw data is included and no post-screening process is applied, because we want to simulate the real crowdsourcing scenario where for the initial several rounds of the experiment we do not have enough data to evaluate the rater’s reliability.

At each round of the experiment 120 observations were obtained using four different methods:

**HY-ACT:** Our hybrid system with active sampling (described in Section 4.4) and  $\alpha = 0.2$ .

**HY-RND:** Our hybrid system with random sampling – For image  $A_i$ , with probability 0.5 that the MOS test  $\mathcal{E}_i$  is sampled, and with probability 0.5 a PC test is sampled and it is uniformly randomly chosen from  $\{\mathcal{E}_{ij} | j = 1, \dots, n, j \neq i\}$ ;

**MOS:** A standard MOS test, where at each iteration of the experiment, we make one additional MOS observation for each image.

**HRRG** [68]: A standard PC test with random sampling. At each iteration, 120 random pairs are sampled based on Erdős-Rényi random graphs.

After each iteration of the experiment, estimates of the underlying scores are obtained using all previously observed data. In particular, HY-ACT and HY-RND estimate the underlying scores by solving the MAP problem described in Section 4.3.3. MOS simply takes the average of all observations for one particular image as its score. HRRG uses the HodgeRank [26] with an angular transform model to obtain the underlying score. In the first iteration of the experiment, HY-ACT and HY-RND were initialized with 120 MOS tests. After initialization, 150 rounds of experiments were performed and in the end a total of  $151 \times 120 = 18120$  observations

were obtained. In this experiment, we simply set  $\gamma = \{1, 2, 3\}$  since we found that with this approximation the performance does not vary much and it is faster to run the experiment. The process was repeated 100 times and the median values of the Kendall's  $\tau$  correlation and LCC are presented in Figs. 4.5 and 4.6, where the x-axis represents the number of observations for all 120 images. When the number of observations is very small (for example, less than 10 observations can be made for each image), the HY-ACT curve and the MOS curve are almost identical. This is because in the first several rounds of the experiment, MOS tests have higher EIGs than PC tests and all 120 selected tests are MOS tests. As more observations are obtained, the active sampling method starts to sample more PC tests. Fig. 4.7 shows the average number of MOS and PC tests performed at each iteration of the experiment.

Due to high observation noise associated with the MOS test, the Kendall's  $\tau$  coefficients of the MOS test is low even with a large number of rating observations and the PC test is indeed more accurate than the MOS test. The active sampling is critical to the success of the hybrid test, since when a random sampling method as in HY-RND is used, the performance of the hybrid system drops. Table 4.2 shows the average number of observations required for each image to achieve a given Kendall's  $\tau$  coefficient. Compared to HRRG, HY-ACT significantly reduced the required number of observations to achieve a given accuracy. Fig. 4.8 shows the standard deviation (STD) of the Kendall's  $\tau$  coefficients of the 100 repeated experiments. We can see that HY-ACT has smaller STD than other methods, which implies that HY-ACT has more consistent performance and is thus more reliable.

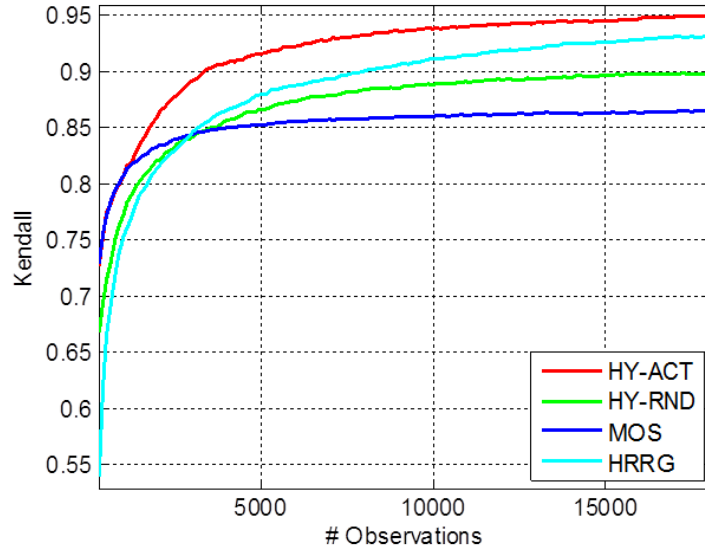


Figure 4.5: Kendall's  $\tau$  in the Crowdsourcing experiment.

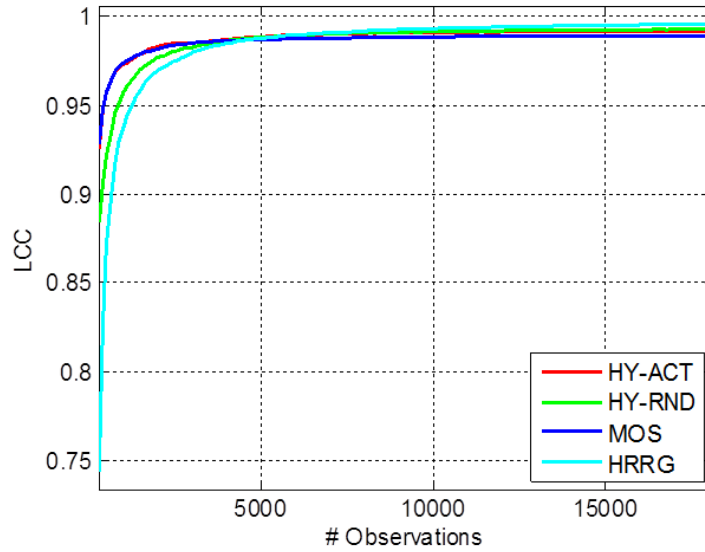


Figure 4.6: LCC in the Crowdsourcing experiment.

## 4.6 Discussion and Open Problems

### 4.6.1 Discussion

We have introduced a hybrid system for subjective IQA. Our system combines MOS and PC tests via a unified probabilistic model for estimating the underlying

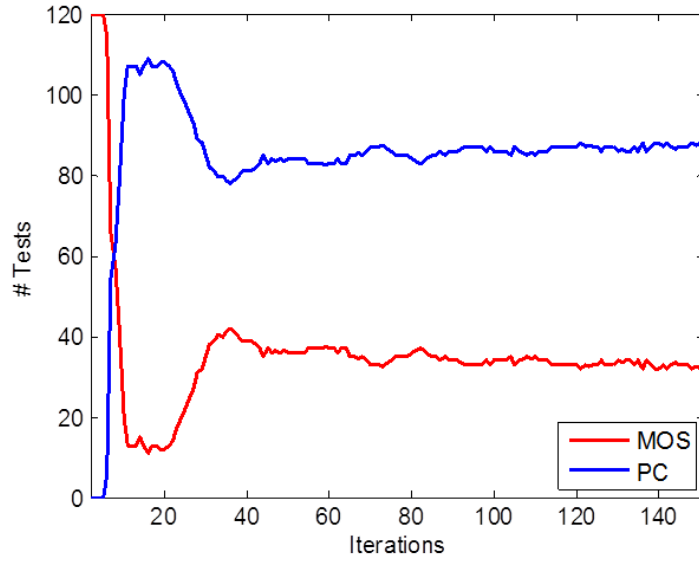


Figure 4.7: Number of MOS and PC tests sampled in each iteration.

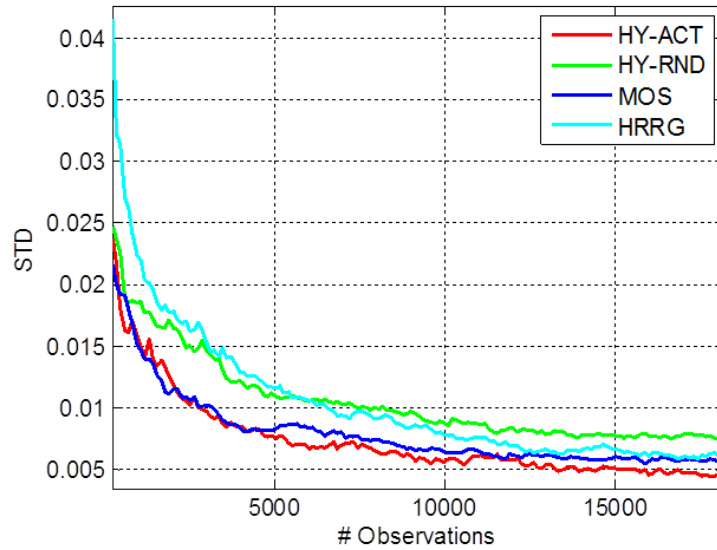


Figure 4.8: Standard deviation of Kendall's  $\tau$ .

quality scores of images. An active sampling method has been introduced to efficiently construct queries of tests which maximize the expected information gain. This method assumes that the MOS test is less accurate than the PC test, which is usually true for subjective IQA. However, when highly accurate MOS test is avail-

Kendall	0.85	0.90	0.91	0.92	0.93
HRRG	27	67	82	107	138
HY-ACT	15	28	36	47	61

Table 4.2: Average number of required observations per image for achieving a given Kendall’s  $\tau$ .

able, the experimenter may prefer the MOS test for subjective IQA. This is because when similar observation noise is involved in the MOS test and the PC test, MOS tests are more informative than PC tests.

## 4.6.2 Open Problems

There are a number of open problems and our hybrid IQA design can be extended in the following ways:

### 1. Non-constant noise variances

Our current model assumes that the variances of observation noise for different images in different types of tests are the same. This assumption does not necessarily hold in practice. By taking into consideration these factors, one may further improve our current model. One way for achieving this is to assume different noise variances for MOS and PC tests. In our experiments, we have empirically increased the noise variances for the MOS test to be slightly higher than that of the PC test. If training data is available, the relationship between these two noise variances can be learned in a more systematical way.

### 2. Online Learning

In the current setting, when new observations are obtained, our model has to be

retrained, i.e. the MAP problem has to be solved again using all available observations. This process can be time-consuming when the number of objects is large. A more efficient way of updating the model is to use online learning so that one can update the posterior probability using only the new observations. In particular, one can achieve this by using the techniques introduced in [83].

### **3. Incorporate Image Features**

In our experiments, we have used an uninformative prior for the underlying score, however, when additional information about the underlying score is available, it can easily be incorporated into our model by constructing the prior distribution using prior information. In particular, one may use image features associated with each image to form the covariance matrix for the prior distribution.

### **4. Improving MOS and PC tests**

Subjective IQA is still an active research field. Different designs for MOS and PC tests have been proposed. For example, in the DSIS (double-stimulus impairment scale) test [84], each test unit presents a reference and its distorted version so that the subject can evaluate how much degradation has been introduced. For the PC test, one may allow ties by allowing the subject to choose “no preference”. The current hybrid test incorporates the MOS test and the PC test in their simplest forms. MOS tests and PC tests in more complicated forms as mentioned above can be easily incorporated in our framework.



## Chapter 5: Conclusions

In this thesis, we have defined and addressed three important problems related to image quality assessment: general-purpose NR-IQA, “opinion-free” NR-IQA and subjective IQA.

### 5.1 Feature Learning for No-Reference Image Quality Assessment

In Chapter 2, we addressed the problem of general-purpose No-Reference Image Quality Assessment (NR-IQA), where the objective is to build computational models that can automatically predict the quality of digital images without access to the non-distorted reference image and without prior knowledge of the types of distortions. Our specific contributions include:

- The introduction of CORNIA – a feature learning framework for general-purpose NR-IQA.
- An effective approach to learn IQA feature – CORNIA learns features (unsupervised or supervised) for discriminating image quality degradation levels and yields state-of-the-art performance on standard IQA benchmarks.
- A real-time NR-IQA framework – Feature computation is extremely fast for

CORNIA and therefore it can be used in real-time applications.

- Domain Adaptation – Unlike previous methods which are usually limited to natural scene image domain, CORNIA can be applied to different image domains.

## 5.2 “Opinion-free” No-Reference Image Quality Assessment

In Chapter 3, we addressed the problem of learning NR-IQA models without human opinion scores. Obtaining human opinion scores can be a time-consuming and expensive process, it is therefore desirable to train a quality prediction model without using human opinion scores. Our specific contributions include:

- The introduction of an unsupervised FR measure combination method – we apply unsupervised rank aggregation to combine multiple FR measures into a single synthetic score. The combined synthetic score outperforms each individual FR measure.
- An effective way for extending existing “opinion-aware” NR-IQA models to “opinion-free” NR-IQA models – We use the combined synthetic score or a single FR measure to replace the human opinion score in training NR-IQA model. In both cases, the results are obtained at significantly reduced cost. The NR-IQA models trained on synthetic scores are comparable to models trained on human opinion score and significantly outperform previous “opinion-free” NR-IQA models.

### 5.3 Active Learning for Subjective Image Quality Assessment

In Chapter 4, we addressed the problem of subjective IQA, where the objective is to obtain gold standard labels for image quality score based on humans' subjective judgments. Our specific contributions include

- A hybrid system which combines the MOS test and the PC test via a unified probabilistic model for estimating the underlying quality scores of images.
- An active sampling method which efficiently constructs queries of tests that maximize the expected information gain. Our method effectively reduces the required number of observations for achieving a certain accuracy and improves on the state-of-the-art.

### 5.4 Publications

#### 5.4.1 Feature Learning for NR-IQA

1. P. Ye and D. Doermann, “No-reference Image Quality Assessment based on Visual Codebook,” in IEEE International Conference on Image Processing (ICIP), pp. 3150-3153, 2011.
2. P. Ye and D. Doermann, “No-Reference Image Quality Assessment using Visual Codebooks”, IEEE Trans. on Image Processing, vol.21, no.7, pp.3129-3138, 2012.
3. P. Ye, J. Kumar, L. Kang and D. Doermann, “Unsupervised Feature Learning

Framework for No-reference Image Quality Assessment”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1098-1105, 2012.

4. P. Ye and D. Doermann. “Learning Features for Predicting OCR Accuracy.” International Conference on Pattern Recognition (ICPR), pp. 3204–3207, 2012.
5. P. Ye, J. Kumar, L. Kang and D. Doermann. “Real-time No-Reference Image Quality Assessment based on Filter Learning.” International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 987–994, 2013.
6. P. Ye and D. Doermann. “Document Image Quality Assessment: A Brief Survey.” International Conference on Document Analysis and Recognition (ICDAR), pp. 723–727, 2013.
7. P. Ye, J. Kumar and D. Deormann. “Feature Learning for No-Reference Image Quality Assessment” IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI). (Submitted)

#### 5.4.2 “Opinion-free” NR-IQA

P. Ye and D. Doermann. “Beyond Human Opinion Scores: Blind Image Quality Assessment based on Synthetic Scores” CVPR2014 (Submitted).

### 5.4.3 Active Learning for subjective IQA

1. P. Ye and D. Doermann. “Combining preference and absolute judgements in a crowd-sourced setting.”, June 2013. (ICML’13 workshop: Machine Learning Meets Crowdsourcing).
2. P. Ye and D. Doermann. “Active Sampling for Subjective Image Quality Assessment” CVPR2014 (Submitted).

## Appendix A

Evidence maximization in Eq. 4.11 is equivalent to find the minimizer of the negative logarithm of the evidence:

$$-\log(\Pr(P, M|\gamma)) \approx \mathcal{F}(\hat{s}) + \frac{1}{2}\log|\hat{R}| + \text{const} \quad (1)$$

where  $\hat{s} = \text{argmax}_s \mathcal{F}(s)$  and  $\hat{R} = R|_{s=\hat{s}}$  and

$$R = \frac{\partial^2 \mathcal{F}(s)}{\partial s \partial s^T} = -\frac{\partial^2 \log(\Pr(P|s))}{\partial s \partial s^T} - \frac{\partial^2 \log(\Pr(M|s, \gamma))}{\partial s \partial s^T} + \Omega^{-1} \quad (2)$$

Let's denote  $R_p \equiv \frac{\partial^2 \log(\Pr(P|s))}{\partial s \partial s^T}$  and  $R_m \equiv \frac{\partial^2 \log(\Pr(M|s, \gamma))}{\partial s \partial s^T}$  and let  $\hat{R}_p$  and  $\hat{R}_m$

denote the  $R_p$  and  $R_m$  at the MAP estimate.

The negative logarithm of the evidence can then be written as follows:

$$\begin{aligned} -\log(\Pr(P, M|\gamma)) &\approx -\log(\Pr(P|\hat{s})) - \log(\Pr(M|\hat{s}, \gamma)) + \frac{1}{2}(\hat{s} - \mu)^T \Omega^{-1}(\hat{s} - \mu) \\ &\quad + \frac{1}{2}\log|\hat{R}_p - \hat{R}_m + \Omega^{-1}| + \text{const} \end{aligned} \quad (3)$$

$\gamma = [\gamma_1, \gamma_2, \gamma_3, \gamma_4]$  are cutoff parameters in MOS observation model. Since  $\gamma_1 < \gamma_2 < \gamma_3 < \gamma_4$ , we may assume  $\gamma_{k+1} - \gamma_k = \Delta_{k+1}$ , where  $\Delta_k > 0$  and  $k = 2, 3, 4$ . We next compute gradient of Eq. 3 with respect to  $\gamma = [\gamma_1, \Delta_2, \Delta_3, \Delta_4]$ . This definition of parameters is helpful to convert the constrained optimization problem into an unconstrained optimization problem. To solve this optimization problem,

we need to compute the partial derivatives of  $\partial - \log Pr(D|\gamma)/\partial\gamma$ .

In Eq. 3,  $\log(Pr(M|\hat{s}, \gamma))$  and  $\hat{R}_m$  are explicit functions of  $\gamma$ . Since  $\hat{s}$  depends on  $\gamma$ , other terms in Eq. 3 are implicit functions of  $\gamma$ . The partial derivative of Eq. 3 w.r.t  $\gamma$  can be written as:

$$\frac{\partial - \log Pr(P, M|\gamma)}{\partial\gamma} = -\frac{\partial \log Pr(P, M|\gamma)}{\partial\gamma}\bigg|_{explicit} - \sum_{i=1}^n \frac{\partial \log(Pr(P, M|\gamma))}{\partial \hat{s}_i} \frac{\partial \hat{s}_i}{\partial\gamma} \quad (4)$$

Without loss of generality, in the following of this section, we assume  $\sigma = 1/\sqrt{2}$ .

## A.1 Explicit Differentiation

$$-\frac{\partial \log Pr(P, M|\gamma)}{\partial\gamma}\bigg|_{explicit} = -\frac{\partial \log Pr(M|\hat{s}, \gamma)}{\partial\gamma} + \frac{1}{2} \frac{\partial \log |-\hat{R}_p - \hat{R}_m + \Omega^{-1}|}{\partial\gamma} \quad (5)$$

Assume  $z_k^i \equiv \frac{\gamma_k - s_i}{\sigma} = \sqrt{2}(\gamma_k - s_i)$ .

$$\frac{\partial - \log(Pr(M|s, \gamma))}{\partial\gamma_1} = -\sqrt{2} \sum_{i=1}^n (M_{i,1} \frac{\phi(z_1^i)}{\Phi(z_1^i) - \Phi(z_0^i)} + \sum_{k=2}^5 M_{i,k} \frac{\phi(z_k^i) - \phi(z_{k-1}^i)}{\Phi(z_k^i) - \Phi(z_{k-1}^i)}) \quad (6)$$

$$\frac{\partial -\log(\Pr(M|s, \gamma))}{\partial \Delta_j} = -\sqrt{2} \sum_{i=1}^n (M_{i,j} \frac{\phi(z_j^i)}{\Phi(z_j^i) - \Phi(z_{j-1}^i)} + \sum_{k=j+1}^5 M_{i,k} \frac{\phi(z_k^i) - \phi(z_{k-1}^i)}{\Phi(z_k^i) - \Phi(z_{k-1}^i)}) \quad (7)$$

$$\frac{\partial \log|-\hat{R}_p - \hat{R}_m + \Omega^{-1}|}{\partial \gamma} = \text{Trace}((\hat{R}_p + \hat{R}_m - \Omega^{-1})^{-1} \frac{\partial \hat{R}_m}{\partial \gamma}) \quad (8)$$

## A.2 Implicit Differentiation

$\hat{s}$  is the minimum of  $\mathcal{F}(s)$ , therefore we have  $\partial \mathcal{F}(s)/\partial s = 0$  at  $s = \hat{s}$  and the implicit derivatives of the first three terms of Eq. 3 vanish, leaving only

$$\frac{\partial -\log(\Pr(P, M|\gamma))}{\partial \hat{s}_i} \frac{\partial \hat{s}_i}{\partial \gamma} = \frac{1}{2} \frac{\partial \log|\hat{R}|}{\partial \hat{s}_i} \frac{\partial \hat{s}_i}{\partial \gamma} \quad (9)$$

**Compute  $\partial|\hat{R}|/\partial \hat{s}$**

$$\begin{aligned} \frac{\partial \log|\hat{R}|}{\partial s_i} &= \text{Trace}(\hat{R}^{-1} \frac{\partial \hat{R}}{\partial s_i}) \\ &= \text{Trace}(\hat{R}^{-1} (-\frac{\partial \hat{R}_p}{\partial s_i} - \frac{\partial \hat{R}_m}{\partial s_i} + \frac{\partial \Omega^{-1}}{\partial s_i})) \\ &= \text{Trace}(\hat{R}^{-1} (-\frac{\partial \hat{R}_p}{\partial s_i} - \frac{\partial \hat{R}_m}{\partial s_i})) \end{aligned} \quad (10)$$

**Compute  $\partial \hat{s}/\partial \gamma$**

$\partial \hat{s}/\partial \gamma$  is determined by the following function implicitly:

$$\frac{\partial \mathcal{F}(s)}{\partial s} \Big|_{s=\hat{s}} = 0 \quad (11)$$



$$\begin{aligned}
& \frac{\partial}{\partial \gamma_1} \left( \frac{\partial \mathcal{F}(s)}{\partial s} \right) + \frac{\partial^2 \mathcal{F}(s)}{\partial s \partial s^T} \frac{\partial s}{\partial \gamma_1} \Big|_{s=\hat{s}} = 0 \\
& \Rightarrow \hat{R} \frac{\partial \hat{s}}{\partial \gamma_1} = \frac{\partial}{\partial \gamma_1} \frac{\partial \log Pr(M|s, \gamma)}{\partial s} \\
& \Rightarrow \frac{\partial \hat{s}}{\partial \gamma_1} = \hat{R}^{-1} \Psi_{\gamma_1}
\end{aligned} \tag{12}$$

Similarly, we have

$$\frac{\partial \hat{s}}{\partial \Delta_k} = \hat{R}^{-1} \Psi_{\Delta_k} \tag{13}$$

where  $k = 2, 3, 4$  and  $\Psi_{\Delta_k} = \frac{\partial}{\partial \Delta_k} \frac{\partial \log Pr(M|s, \gamma)}{\partial s}$ .

### A.3 Computation Details

For ease of representation, we define:

$$v_l(k, i) = \frac{(z_k^i)^l \phi(z_k^i) - (z_{k-1}^i)^l \phi(z_{k-1}^i)}{\Phi(z_k^i) - \Phi(z_{k-1}^i)}$$

$$u_l(k, i) = \frac{(z_k^i)^l \phi(z_k^i)}{\Phi(z_k^i) - \Phi(z_{k-1}^i)}$$

$$w(i, j) = \frac{\phi(s_i - s_j)}{\Phi(s_i - s_j)}$$

### A.3.1 Derivative of $R_m$

$R_m$  is a diagonal matrix and its diagonal element is determined by:

$$R_m(i, i) = \frac{\partial^2 L(s|M)}{\partial s_i^2} = -2 \sum_{k=1}^5 M_{i,k} \left[ \left( \frac{\phi(z_k^i) - \phi(z_{k-1}^i)}{\Phi(z_k^i) - \Phi(z_{k-1}^i)} \right)^2 + \frac{z_k^i \phi(z_k^i) - z_{k-1}^i \phi(z_{k-1}^i)}{\Phi(z_k^i) - \Phi(z_{k-1}^i)} \right] \quad (14)$$

$$\frac{\partial R_m(i, i)}{\partial s_i} = -2\sqrt{2} \sum_{k=1}^5 M_{i,k} (2v_0(k, i)^3 + 3v_0(k, i)v_1(k, i) - v_0(k, i) + v_2(k, i)) \quad (15)$$

and

$$\frac{\partial R_m(i, i)}{\partial s_j} = 0 \text{ For } j \neq i$$

$$\begin{aligned} \frac{\partial R_m(i, i)}{\partial \gamma_1} &= 2\sqrt{2} [M_{i,1} (2v_0(1, i)^2 u_0(1, i) + 2v_0(1, i)u_1(1, i) + v_1(1, i)u_0(1, i) - u_0(1, i) \\ &\quad + u_2(1, i)) + \sum_{k=2}^5 M_{i,k} (2v_0(k, i)^3 + 3v_0(k, i)v_1(k, i) - v_0(k, i) + v_2(k, i))] \end{aligned} \quad (16)$$

$$\begin{aligned} \frac{\partial R_m(i, i)}{\partial \Delta_j} &= 2\sqrt{2} [M_{i,j} (2v_0(j, i)^2 u_0(j, i) + 2v_0(j, i)u_1(j, i) + v_1(j, i)u_0(j, i) - u_0(j, i) \\ &\quad + u_2(j, i)) + \sum_{k=j+1}^5 M_{i,k} (2v_0(k, i)^3 + 3v_0(k, i)v_1(k, i) - v_0(k, i) + v_2(k, i))] \end{aligned} \quad (17)$$

### A.3.2 Derivative of $R_p$

$R_p$  is a  $n$ -by- $n$  matrix determined by:

$$\begin{aligned} R_p(i, j) &= P_{i,j} \frac{\phi(s_i - s_j)}{\Phi(s_i - s_j)} (s_i - s_j + \frac{\phi(s_i - s_j)}{\Phi(s_i - s_j)}) + P_{j,i} \frac{\phi(s_j - s_i)}{\Phi(s_j - s_i)} (s_j - s_i + \frac{\phi(s_j - s_i)}{\Phi(s_j - s_i)}) \\ &= P_{i,j} w(i, j) (s_i - s_j + w(i, j)) + P_{j,i} w(j, i) (s_j - s_i + w(j, i)) \end{aligned} \quad (18)$$

where  $w(i, j) = \frac{\phi(s_i - s_j)}{\Phi(s_i - s_j)}$

$$R_p(i, i) = - \sum_{j \neq i} R_p(i, j) \quad (19)$$

$$\frac{\partial w(i, j)}{\partial s_i} = - \frac{\phi(s_i - s_j)}{\Phi(s_i - s_j)} \left( \frac{\phi(s_i - s_j)}{\Phi(s_i - s_j)} + s_i - s_j \right) = -w(i, j)(w(i, j) + s_i - s_j) \quad (20)$$

$$\frac{\partial w(j, i)}{\partial s_i} = w(j, i)(w(j, i) + s_j - s_i) \quad (21)$$

$$\begin{aligned} \frac{\partial R_p(i, j)}{\partial s_i} &= P_{i,j} [(s_i - s_j + w(i, j)) \frac{\partial w(i, j)}{\partial s_i} + w(i, j)(1 + \frac{\partial w(i, j)}{\partial s_i})] \\ &\quad + P_{j,i} [(s_j - s_i + w(j, i)) \frac{\partial w(j, i)}{\partial s_i} + w(j, i)(-1 + \frac{\partial w(j, i)}{\partial s_i})] \\ &= P_{i,j} \frac{\partial w(i, j)}{\partial s_i} (s_i - s_j + 2w(i, j)) + P_{i,j} w(i, j) \\ &\quad + P_{j,i} \frac{\partial w(j, i)}{\partial s_i} (s_j - s_i + 2w(j, i)) - P_{j,i} w(j, i) \end{aligned} \quad (22)$$

$$\frac{\partial R_p(i, j)}{\partial s_j} = - \frac{\partial R_p(i, j)}{\partial s_i}$$

$$\frac{\partial R_p(i, j)}{\partial s_k} = 0 \text{ for } k \neq i, k \neq j$$

### A.3.3 Derivative of $\partial \log Pr(M|s, \gamma)/\partial s$

$$\begin{aligned} \Psi_{\Delta_k}(i) &= \frac{\partial}{\partial \Delta_k} \frac{\partial \log Pr(M|s, \gamma)}{\partial s_i} \\ &= -\frac{\partial}{\partial \Delta_k} \frac{1}{\sigma} \sum_{j=1}^5 M_{i,j} \frac{\phi(z_j^i) - \phi(z_{j-1}^i)}{\Phi(z_j^i) - \Phi(z_{j-1}^i)} \\ &= -\frac{1}{\sigma} \sum_{j=1}^5 M_{i,j} \left\{ -\frac{\phi(z_j^i) \frac{\partial z_j^i}{\partial \Delta_k} - \phi(z_{j-1}^i) \frac{\partial z_{j-1}^i}{\partial \Delta_k}}{(\Phi(z_j^i) - \Phi(z_{j-1}^i))^2} (\phi(z_j^i) - \phi(z_{j-1}^i)) \right. \\ &\quad \left. - \frac{z_j^i \phi(z_j^i) \frac{\partial z_j^i}{\partial \Delta_k} - z_{j-1}^i \phi(z_{j-1}^i) \frac{\partial z_{j-1}^i}{\partial \Delta_k}}{\Phi(z_j^i) - \Phi(z_{j-1}^i)} \right\} \end{aligned} \quad (23)$$

where for  $j = 2, 3, 4$

$$z_j^i = \sqrt{2}(\gamma_1 + \dots + \Delta_j - s_i)$$

and

$$\frac{\partial z_j^i}{\partial \Delta_k} = \begin{cases} \sqrt{2} & k \leq j \\ 0 & k > j \end{cases} \quad (24)$$

Therefore, we have

$$\Psi_{\Delta_k}(i) = 2M_{i,k}(u_0(k, i)v_0(k, i) + u_1(k, i)) + 2\sum_{j=k+1}^5 M_{i,j}(v_0(j, i)^2 + v_1(j, i))$$

$$\Psi_{\gamma_1}(i) = 2M_{i,1}(u_0(1, i)v_0(1, i) + u_1(1, i)) + 2\sum_{j=2}^5 M_{i,j}(v_0(j, i)^2 + v_1(j, i)) \quad (25)$$

## Bibliography

- [1] Z. Wang and A. C. Bovik. *Modern Image Quality Assessment*. Morgan & Claypool, 2006.
- [2] Peng Ye and David Doermann. Document image quality assessment: A brief survey. In *Intl. Conf. on Document Analysis and Recognition (ICDAR)*, pages 723–727, 2013.
- [3] Rafael Dueire Lins. A taxonomy for noise in images of paper documents - the physical noises. In *Int. Conf. on Image Anal. and Recog.*, pages 844–854, 2009.
- [4] Chaofeng Li, A. C. Bovik, and Xiaojun Wu. Blind image quality assessment using a general regression neural network. *IEEE Transactions on Neural Networks*, 22(5):793–799, may 2011.
- [5] Huixuan Tang, N. Joshi, and A. Kapoor. Learning a blind measure of perceptual image quality. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 305–312, 2011.
- [6] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, Dec. 2011.
- [7] M.A. Saad, A.C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, Aug. 2012.
- [8] A. Mittal, A. Moorthy, and A. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, PP(99):1, 2012.
- [9] Lihuo He, Dacheng Tao, Xuelong Li, and Xinbo Gao. Sparse representation for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1146–1153, 2012.

- [10] A. Souza, M. Cheriet, S. Naoi, and C.Y. Suen. Automatic filter selection using image quality assessment. In *Intl. Conf. on Document Analysis and Recognition (ICDAR)*, pages 508–512, 2003.
- [11] Michael Cannon, Judith Hochberg, and Patrick Kelly. Quality assessment and restoration of typewritten document images. *International Journal on Document Analysis and Recognition (IJDAR)*, 2(2-3):80–89, 1999.
- [12] Deepak Kumar and A. G. Ramakrishnan. QUAD: Quality assessment of documents. In *Int. Workshop on Camera-based Doc. Anal. and Recog.*, pages 79–84, 2011.
- [13] Luis Blando Junichi, Junichi Kanai, and Thomas A. Nartker. Prediction of OCR accuracy using simple image features. In *Intl. Conf. on Document Analysis and Recognition (ICDAR)*, pages 319–322, 1995.
- [14] V. Govindaraju and S. N. Srihari. Image quality and readability. In *Intl. Conf. on Image Processing (ICIP)*, pages 324–327, 1995.
- [15] Xujun Peng, Huaigu Cao, K. Subramanian, R. Prasad, and P. Natarajan. Automated image quality assessment for camera-captured OCR. In *Intl. Conf. on Image Processing (ICIP)*, pages 2621–2624, 2011.
- [16] S.-L. Chou and S.-S. Yu. Sorting qualities of handwritten chinese characters for setting up a research database. In *Intl. Conf. on Document Analysis and Recognition (ICDAR)*, pages 474–477, 1993.
- [17] T. Obafemi-Ajayi and G. Agam. Character-based automated human perception quality assessment in document images. *IEEE Transaction on Systems, Man and Cybernetics, Part A: Systems and Humans*, 42:584–595, 2011.
- [18] Jayant Kumar, Francine Chen, and David Doermann. Sharpness estimation for document and scene images. In *Intl. Conf. on Pattern Recognition (ICPR)*, pages 3292–3295, 2012.
- [19] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised Feature Learning Framework for No-reference Image Quality Assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1098–1105, 2012.
- [20] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Real-time no-reference image quality assessment based on filter learning. In *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 987–994, 2013.
- [21] Peng Ye and David Doermann. Beyond human opinion scores: Blind image quality assessment based on synthetic scores. In *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. submitted.

- [22] Peng Ye and David Doermann. Active sampling for subjective image quality assessment. In *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. submitted.
- [23] Zhou Wang, Alan C. Bovik, and Ligang Lu. Why is image quality assessment so difficult? In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 3313–3316, May 2002.
- [24] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. In *NIPS*, pages 1033–1040, 2008.
- [25] Jianchao Yang, Kai Yu, and T. Huang. Supervised translation-invariant sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3517–3524, 2010.
- [26] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.
- [27] Anil K. Jain and Kalle Karu. Learning texture discrimination masks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(2):195–205, Feb. 1996.
- [28] Huixuan Tang, Neel Joshi, and Ashish Kapoor. Learning a blind measure of perceptual image quality. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 305–312, 2011.
- [29] P. Ye and D. Doermann. No-reference image quality assessment based on visual codebook. In *IEEE Intl. Conf. on Image Processing (ICIP)*, 2011.
- [30] Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.
- [31] Jianchao Yang, Kai Yu, Yihong Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1794–1801, 2009.
- [32] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, T. Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3360–3367, Jun. 2010.
- [33] Adam Coates and Andrew Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *International Conference on Machine Learning (ICML)*, pages 921–928, June 2011.
- [34] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

- [35] Lutz Prechelt. Early stopping - but when? In *Neural Networks: Tricks of the Trade, volume 1524 of LNCS, Chapter 2*, pages 55–69. Springer-Verlag, 1997.
- [36] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. LIVE image quality assessment database release 2. Online, <http://live.ece.utexas.edu/research/quality>.
- [37] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.
- [38] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. Tid2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radio Electronics*, 10:30–45, 2009.
- [39] Eric C. Larson and Damon M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010.
- [40] Dinesh Jayaraman, Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. Objective image quality assessment of multiply distorted images. In *Proceedings of Asilomar Conference on Signals, Systems and Computers*, 2012.
- [41] Jayant Kumar, Peng Ye, and David Doermann. A Dataset for Quality Assessment of Camera Captured Document Images. In *International Workshop on Camera-Based Document Analysis and Recognition (CBDAR)*, pages 39–44, Aug. 2013.
- [42] Isabelle Guyon, Robert M. Haralick, Jonathan J. Hull, and Ihsin Tsaiyun Phillips. Data sets for OCR and document image understanding research. In *In Proceedings of the SPIE - Document Recognition IV*, pages 779–799. World Scientific, 1997.
- [43] D. Lewis. Building a test collection for complex document information processing. In *in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 665–666, 2006.
- [44] ABBYY Finereader 10 Professional Edition, build 10.0.102.74. 2009.
- [45] ISRI-OCR evaluation tool: Code and data to evaluate OCR accuracy, originally from UNLV/ISRI. <http://code.google.com/p/isri-ocr-evaluation-tools/>, January 2010.
- [46] K. Seshadrinathan and A.C. Bovik. Temporal hysteresis model of time varying subjective video quality. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1153–1156, 2011.



- [47] A. Mittal, G. S. Muralidhar, and A. C. Bovik. Blind image quality assessment without human training using latent quality factors. *IEEE Signal Processing Letters*, 19 (2):75–78, 2012.
- [48] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal processing Letters*, 22:209–212, 2013.
- [49] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 995–1002, 2013.
- [50] Tsung-Jung Liu, Weisi Lin, and C.-C. Jay Kuo. Image quality assessment using multi-method fusion. *IEEE Transactions on Image Processing*, 22(5):1793–1807, 2013.
- [51] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *arXiv:1308.3052*, 2013.
- [52] H. R. Sheikh, A. C. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, Dec. 2005.
- [53] Lin Zhang, D. Zhang, Xuanqin Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.
- [54] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, 2011.
- [55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [56] G. V. Cormack, C. L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009.
- [57] Maksims N. Volkovs and Richard S. Zemel. Supervised CRF framework for preference aggregation. In *International Conference on Information and Knowledge Management (CIKM)*, 2013.
- [58] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval Journal*, 13(4):346–374, Aug. 2010.

- [59] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment – Phase II, Aug. 2003. available at <http://www.vqeg.org/>.
- [60] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, pages 1871–1874, 2008.
- [61] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. A crowd-sourceable qoe evaluation framework for multimedia content. In *Proceedings of ACM Multimedia*, 2009.
- [62] W.S. Torgerson. *Theory and methods of scaling*. John Wiley & Sons, New York, 1958.
- [63] Subjective video quality assessment methods for multimedia applications. ITU-T Recommendation P.910, Apr. 2008.
- [64] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or there: preference judgments for relevance. In *Proceedings of the IR research, 30th European Conference on Advances in information retrieval*, pages 16–27, Berlin, Heidelberg, 2008.
- [65] F. Ribeiro, D. Florencio, Cha Zhang, and M. Seltzer. Crowdmoss: An approach for crowdsourcing mean opinion score studies. In *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2416–2419, May 2011.
- [66] L.L. Thurstone. A law of comparative judgement. *Psychological Review*, 1927. 34:273-286.
- [67] H.A. David. *The Method of Paired Comparisons*. Hodder Arnold, second edition, 1988.
- [68] Qianqian Xu, Tingting Jiang, Yuan Yao, Qingming Huang, Bowei Yan, and Weisi Lin. Random partial paired comparison for subjective video quality assessment via hodgerank. In *ACM International Conference on Multimedia*, pages 393–402. ACM, 2011.
- [69] Qianqian Xu, Qingming Huang, and Yuan Yao. Online crowdsourcing subjective image quality assessment. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 359–368. ACM, 2012.
- [70] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pair-wise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 193–202. ACM, 2013.

- [71] Thomas Pfeiffer, Xi Alice Gao, Andrew Mao, Yiling Chen, and David G. Rand. Adaptive polling and information aggregation. In *The 26th Conference on Artificial Intelligence (AAAI)*, 2012.
- [72] Maksims N. Volkovs and Richard S. Zemel. A flexible generative model for preference aggregation. In *Proceedings of the 21st International Conference on World Wide Web*, pages 479–488, 2012.
- [73] Frederick Mosteller. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16:3–9, 1951.
- [74] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- [75] David F. Gleich and Lek-heng Lim. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 60–68, 2011.
- [76] Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, December 2005.
- [77] Adriano Azevedo-Filho and Ross D. Shachter. Laplace’s method approximations for probabilistic inference in belief networks with continuous variables. In *Uncertainty in Artificial Intelligence*, pages 28–36, 1994.
- [78] Jr. Ford, L. R. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8):pp. 28–33, 1957.
- [79] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.
- [80] Mark E. Glickman and Shane T. Jensen. Adaptive paired comparison design. *Journal of Statistical Planning and Inference*, 127:2005.
- [81] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C : The Art of Scientific Computing*. Cambridge University Press, Oct. 1992.
- [82] Andreas Wächter and Lorenz T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106:25–57, 2006.
- [83] Ruby C. Weng and Chih-Jen Lin. A Bayesian approximation method for online ranking. *Journal of Machine Learning Research*, 12:267–300, 2011.
- [84] Methodology for the subjective assessment of the quality of television pictures. ITU-R Recommendation BT.500-12, Sept. 2009.